Systematic Study and Enhancement of an Implicit Solvent Model

Dissertation
zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)
vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von
Andrea Prunotto
aus
Italien

Promotionskomitee
Prof. Dr. Amedeo Caflisch
Prof. Dr. Camillo de Lellis

Zürich 2010

a Marcello, ël pì cit ëd la nidià

SUMMARY

Molecular dynamics simulations are a computational approach to the study of biological molecules. The treatement of aqueous solvents is a crucial issue to correctly reproduce the functionality and the structure of proteins. The most rigorous treatement of the water solvent consists in the simulations of the solute together with all solvent molecules, including water and ions. However, this approach is highly time consuming (inefficient). One way to speed up the calculation is to replace the single solvent molecules by a potential of mean force that takes into account the mean effect of water molecules on the solute. Several implicit solvent models were hitherto developed, but the balance between accuracy and efficiency is not always satisfactory. This thesis presents the effort to improve the implicit solvent model called FACTS, developed in the Caflisch group since 2003. The main idea of this model is to exploit the *local geometric properties* of groups of solute atoms to rapidly (and analytically) recover the global solvation properties of the solute itself. The result of this work is that the accuracy of the nonpolar treatment of solvation could be significantly enhanced (without any increase in computation time) by introducing a correction (FACTS SISI) based on the Tolman theory of surface tension.

ZUSAMMENFASSUNG

Die Dynamik von Proteinen wird von Computersimulationen nur dann korrekt vorhergesagt, wenn die Atome des Lösungsmittel in das Modell mit aufgenommen werden. Ein Nachteil dieses Vorgehens ist, dass die Simulationen einen immensen Rechenaufwand erfordern. Dieser Bedarf lässt sich drastisch reduzieren, wenn die Granularität des Wassers vernachlässigt wird. Stattdessen wird der komplexe Einfluss des Wassers auf ein Protein durch eine analytische Funktion beschrieben, die die Eigenschaften des Wassermoleküls nur implizit behandelt (daher der englische Fachausdruck *implicit solvent model*). Trotz der Vielzahl der veröffentlichten *implicit solvent models*, ergeben nur wenige Modelle einen guten Kompromis zwischen Genauigkeit und Rechenzeit. In der vorliegenden Arbeiten werden Erweiterungen des erfolgreichen Modells FACTS beschrieben. FACTS basiert auf der Hauptidee, dass lokale und geometrische Eigenschaften des Proteinatomes verwendet werden, um analytisch und (damit) schnell die Lösungsenthalpie zu bestimmen. Im Vergleich zu bisherigen Arbeiten konnte die Genauigkeit der Vorhersage von Lösungsenergie signifikant verbessert werden. Diese Verbesserung basiert auf die Einführung eines Korrekturterms (FACTS SISI), der auf Tolmans Theorie der gekrümmten Flächen basiert. Für die Anwendung von entscheidender Wichtigkeit ist, dass die FACTS SISI Korrektur den Berechnungsaufwand nicht verändert.

Contents

Chapter 1

Introduction

This is a general overview about the problem of simulating biomolecules behaviour with the computer. Physics principles and basic algorithms "behind the scene" are here briefly introduced.

Fere libenter homines id quod volunt credunt.

C. J. Cæsar

The subject of this doctoral thesis is twofold: on one hand, it resumes the outcomes of an extensive *systematic study* of the implicit solvent model called FACTS (Fast Analytical Continuum Treatment of Solvation)[1], an approach that speeds up computer simulations of biological molecules in the framework of the CHARMM forcefield; on the other hand, it outlines *improvements* to this model (in particular, in the nonpolar contribution to solvation energy), motivated by a tendency to instability of the original version of FACTS with proteins.

Chapters 1 and 2 are devoted to a brief introduction to essential concepts of molecular dynamics simulations and to the problem of solvation from a computational point of view. Chapter 3 will introduce the reader to the basic insights of FACTS. Chapter 4 and 5 are devoted to the aforementioned systematic study of FACTS and to the error analysis of FACTS parametrisation. The last chapter contains the theoretical basis and the results of a correction to the original version (based on the microscopic studies of *surface tension* in liquids) which enhances the reliability of FACTS without increase in computation time.

[1] Haberthür et al., *FACTS: Fast Analytical Continuum Treatment of Solvation*, JCC, 29, 701-715 (2008).

1.1 Molecular dynamics simulations

Computer simulations in biochemistry are performed to understand the properties of assemblies of molecules, either in terms of their structure or in terms of their microscopic interactions. They should be seen as a complement to traditional experiments, a new way to make hypothesis and suggest (in the researcher's work) new experiments. There exist two main kinds of simulation techniques: molecular dynamics (MD) and Monte Carlo (MC) simulations. Both exploit the available computing technology: the former to integrate the equation of motions related to a physical system, whereas the latter rely on repeated random sampling to compute measures of physical quantities of the system. Moreover, there is a range of hybrid techniques applying features from both of the methods. This thesis is fully based on MD simulations. The advantage of MD over MC simulations is that they allow to explore the *dynamical* properties of the system, since they calculate the evolution of the studied system in time.

MD computer simulations can be seen as a link between microscopic length and time scales and the macroscopic world of the biochemistry laboratory: they provide insight about the interactions between molecules at an atomic level. The MD predictions can be performed precisely *ad libitum* (provided a deep knowledge about the inherent physics and our capability of integrate non linear equations): limitations are actually imposed by the computer budget, either in terms of hardware cost and/or in terms of computation time. Simulations can be thought of as link between the micro- and macroscopical world and also between theory and experiment. It is possible to test a theory by running a simulation using the same biochemical system: then, the results related to this system can be tested by comparing *in silico* and experimental results, allowing us to discriminate between theories.

MD simulations consist of numerical, step-by-step solutions to the classical equations of motion for a particle i with mass m_i at position $\vec{r}_i$ and subjected to a resultant force $\vec{f}_i$. For a simple atomic system, it can be written as

$$\vec{f}_i = m_i \cdot \ddot{\vec{r}}_i \qquad \vec{f}_i = -\frac{dU}{dr_i}. \tag{1.1}$$

It is necessary to calculate the forces $\vec{f}_i$ acting on the i atoms, and these are usually derived from a potential energy $U(r_i)$, where $i = 1, 2 \ldots N$ represents the complete

set of $3 \cdot N$ atomic coordinates. In this section we focus on this function $U(r_i)$. The part of the potential energy U^{nb} representing non-bonded interactions between atoms is usually split into 1-body, 2-body, 3-body ... terms:

$$U^{nb} = \sum_{i}^{N} u(\vec{r}_i) + \sum_{i} \sum_{j<i} v(\vec{r}_i; \vec{r}_j) + \dots \tag{1.2}$$

where the $u(\vec{r}_i)$ terms represent externally applied potential fields or the effects of the container walls – but they are normally dropped for periodic simulations of bulk systems. Moreover, usually only the pair potential $v(\vec{r}_i; \vec{r}_j) = v(r_{ij})$ is taken into account, while the three-body (and higher order) interactions are neglected[2]. This chapter will be devoted to continuous, differentiable pair-potentials, through we note that discontinuous potentials (such as hard spheres potentials) could also be of relevance[3]. The Lennard-Jones 12-6 potential is the most commonly used form:

$$v^{LJ}(r) = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right] \tag{1.3}$$

with two parameters: σ, the atomic diameter, and ϵ, the minimum of the potential (r should taken to be the interatomic distance). This function has been developed in the earliest studies of the properties of liquid Argon[4]. When attractive interactions are of less importance than the excluded-volume effects coming from molecular packing, the potential can be cut off near its minimum, and then shifted upwards[5]. If electrostatic charges are present (and this is the case for most of the active biomolecules), the Coulomb potential has to be added:

$$v^{Coulomb}(r) = \frac{q_1 \cdot q_2}{4\pi\epsilon_0 r} \tag{1.4}$$

where q_1 and q_2 are the charges and ϵ_0 is the vacuum permittivity. For molecular systems, molecules are simply build out of site-site potentials of the form of Eq. 1.3, 1.4. Quantum calculations can be used to estimate the electron density throughout the

[2] Maitland et al. *Intermolecular forces: their origin and determination.* Clarendon Press, Oxford (1981); Gray et al. *Theory of molecular fluids.* Clarendon Press, Oxford (1984); Sprik *Effective pair potentials and beyond*, Michael Allen and Tildesley editors, Dordrecht (1993); Stone *The Theory of Intermolecular Forces.* Clarendon Press, Oxford (1996).

[3] Allen et al. *Hard convex body fluids*, Adv. Chem. Phys., 86, 1-166 (1993).

[4] Rahman *Correlations in the motion of atoms in liquid argon.* Phys. Rev. A, 136, 405-411 (1964).

[5] Weeks et al. *Role of repulsive forces in determining the equilibrium structure of simple liquids.* J. Chem. Phys., 54, 5237-5247 (1971).

biomolecule. This electron density can be approximated by a distribution of partial charges via Eq. 1.4 – or, more accurately, by a distribution of electrostatic multipoles[6]. It is of of vital importance to consider also the *intramolecular* bonding interactions U^{int}

$$U^{int} = \frac{1}{2}\sum_{bonds} k^r_{i,j}(r_{ij} - r_{eq})^2 + \dots$$

$$+\frac{1}{2}\sum_{bend\ \ angles} k^\theta_{ijk}(\theta_{ijk} - \theta_{eq})^2 + \dots$$

$$+\frac{1}{2}\sum_{torsion\ \ angles}\sum_{m} k^{\phi,m}_{ijkl}(1 + \cos(m\phi_{ijkl} - \gamma_m)). \tag{1.5}$$

The **bonds** involve the distance $r_{ij} = |\vec{r}_i - \vec{r}_j|$ among atomic pairs (a harmonic form is assumed, with a specific equilibrium separation r_{eq}). The **bend angles** θ_{ijk} are defined between bond vectors – such as $|\vec{r}_i - \vec{r}_j|$ and $|\vec{r}_j - \vec{r}_k|$ – and, thus, involve 3 atom coordinates, since $\cos(m\phi_{ijkl}) = -\hat{n}_{ijk} \cdot \hat{n}_{jkl}$, where $\vec{n}_{ijk} = \vec{r}_{ij} \times \vec{r}_{jk}$, $\vec{n}_{jkl} = \vec{r}_{jk} \times \vec{r}_{kl}$ and $\hat{n} = \vec{n}/n$ is the unit normal to the plane defined by each bond pairs. Eventually, **Torsion angles** are written as an expansion (of order m) in term of periodic functions.

1.1.1 Forcefields

The actual potential energy function used in MD simulation is called the **forcefield**. It specifies the precise form of Eq. 1.5, the various parameters k and other relevant constants, the topology and connectivity of the (bio)molecules, the Lennard-Jones parameters and partial charges. This energy function is generally composed by the superposition of two terms U^{int} and U^{nb}:

$$U^{tot} = U^{int} + U^{nb} \tag{1.6}$$

Quantum mechanical calculations may serve as a guide to the best molecular forcefield optimisation. Comparison between simulation results and thermophysical properties (e.g. vibration frequencies) also represents an important tool in forcefield refinement.

[6]Price *Toward more accurate model intermolecular potentials for organic molecules.* Rev. Comput. Chem.,14, 225-289 (2000).

A popular, large-system oriented class of force fields includes AMBER[7], CHARMM[8] and OPLS[9]: they are designed to simulate proteins and other biopolymers (DNA, RNA) in condensed phases: their functional forms are similar to Eq. 1.6.

Once the potential energy function $U(r)$ is given, the next step is to calculate the atomic forces $\vec{f_i} = -\frac{dU}{dr_i}$. At this stage, a theoretical problems arises. The analysis of MD simulations indicates that the motion of a biomolecular system is **chaotic**.

The emergence of chaos is due to an intrinsic sensitivity to initial conditions, which lets similar conformations (at the first stages of an MD simulation) evolve into broadly different structures. The chaotic properties can be identified by nonzero Lyapunov exponents, broad-band power spectra, and strange attractors[10]. The dominant reasons of chaos are mainly the *nonlinear interactions* present in the forcefield[11], the presence of constraints and the stochastic forces generated by the (explicit) solvent.

1.1.2 Coarse-grained forcefields and implicit potentials

The study of long chain molecules is particularly intriguing and led to the adoption of progressively simplified (or "coarse-grained") potential models. Various **implicit** atomic potentials have been devised for the n-alkanes[12]. More approximate potentials have also been designed[13] in which the CH_2 and CH_3 group were represented by single units. These sort of implicit potentials are usually less accurate than the explicit, full-atom potentials, but significantly less expensive (in term of simulation time). Comparisons have been made between the two approaches[14].

[7] Weiner et al. *A new forcefield for molecular mechanical simulation of nucleic acids and proteins.* J. Am. Chem. Soc., 106, 765-784 (1984); Cornell et al. *A 2nd generation forcefield for the simulation of proteins, nucleic-acids, and organic molecules.* J. Am. Chem. Soc., 117, 5179-5197 (1995).

[8] Brooks et al. *CHARMM - A program for macromolecular energy, minimisation, and dynamics calculations.* J. Comput. Chem., 4, 187-217 (1983).

[9] Jorgensen et al. *Development and testing of the OPLS all-atom forcefield on conformational energetics and properties of organic liquids.* J. Am. Chem. Soc.,118, 11225-11236 (1996).

[10] Zhou et al. *Chaos in Biomolecular Dynamics.* J. Phys. Chem., 100, 20, 8101-8105 (1996).

[11] Smith *Chaos a very short introduction* Oxford University Press, 2009.

[12] Chen et al. *Thermodynamic properties of the williams, opls-aa, and mmff94 all-atom forcefields for normal alkanes.* J. Phys. Chem. B, 102, 2578-2586 (1998).

[13] Nath et al. *On the simulation of vapour-liquid equilibria for alkanes.* J. Chem. Phys., 108, 9905-9911 (1998); Martin et al. *Transferable potentials for phase equilibria. 1. united-atom description of n-alkanes.* J. Phys. Chem. B, 102, 2569-2577 (1998).

[14] Tsige et al. *Molecular dynamics simulations and integral equation theory of alkane chains: comparison of explicit and united atom models.* Macromolecules, 36, 2158-2164 (2003).

For more complex biomolecules, the approach has to be improved. In the liquid crystal field, for example, a compromise has been suggested[15]. Using the united atom approach for hydrocarbon chains, but treating phenyl-ring hydrogens explicitly. In polymer simulations it becomes more and more important to further reduce degrees of freedom and, thus, to have more coarse-grained interactions. Significant progress has been made in recent years in approaching this problem[16].

The most fundamental properties of a polymer melt can be captured using a simple chain of pseudo-atoms, together with an attractive infinite nonlinear elastic potential. The key feature of this potential is that it cannot be extended beyond a certain radius, ensuring that the polymer chains cannot move one with respect to the others[17].

Moreover, a phenomenological coarse-grained model of an amphipathic polypeptide, characterised by a free energy profile with distinct amyloid-competent and amyloid-protected states, has been developed[18] and has been used (together with a coarse-grained model of lipid molecules able to self-assemble into bilayer vesicles) to successfully investigate several important features of β-amyloid aggregation.

The aqueous environment strongly influences the thermodynamics and kinetics of all soluble biomolecules. Therefore, it cannot be neglected in modelling applications. On the other hand, high computational costs result from the inclusion of explicit water in simulated systems. Thus, considerable effort has been invested in the development of **implicit solvent models**, in which the influence of the solvent is simply described by a *potential of mean force*.

1.2 Performing MD simulations

1.2.1 Integrating the equations of motion

Consider a system composed of N atoms with coordinates $\vec{r}^N = (\vec{r}_1, \vec{r}_2, \ldots \vec{r}_N)$ and the potential energy $U(\vec{r}^N)$. The atomic momenta $\vec{p}^N = (\vec{p}_1, \vec{p}_2, \ldots \vec{p}_N)$ lead to the kinetic

[15] Garcia et al. *HFF: a forcefield for liquid crystal molecules.* J. Molec. Struc. THEOCHEM, 464, 39-48 (1999).

[16] Reith et al. *CG-OPT: A software package for automatic forcefield design.* Comput. Phys. Commun., 148, 299-313 (2002); Reith et al. *Deriving effective mesoscale potentials from atomistic simulations.* J. Comput. Chem., 24, 1624-1636 (2003).

[17] Pütz et al. What is the entanglement length in a polymer melt? Europhys. Lett., 49, 735-741 (2000).

[18] Pellarin et al. *Interpreting the aggregation kinetics of amyloid peptides.* J Mol Biol. 360, 882-892 (2006).

energy term:

$$K(\vec{p}^N) = \sum_i^N \frac{|\vec{p}_i|^2}{2m_i}.$$

The Hamiltonian can thus be written as $H = K + U$ and the equations of motion become:

$$\dot{\vec{r}}_i = \vec{p}_i/m_i \quad \dot{\vec{p}}_i = \vec{f}_i, \tag{1.7}$$

which is a system of coupled-ordinary differential equations. A great variety of methods has been developed to numerically solve step-by-step the integration of such a system and are implemented in many libraries of the most common programming languages (such as C++, Fortran, python). They are usually the more time-consuming part of a MD simulation. These algorithms must deal either with long or short time scales. There are at least two conflicting features which have to be taken into account during the implementation of these algorithms: first, dynamical properties are captured only if the time-step is not longer than the fastest motion of any substructure of the system at study. Second, a MD run should move on the constant-energy (hyper)surface as long as possible, in order to sample the correct ensemble. On the other hand, to quickly explore the phase space, the time-step has to be set as large as possible, consistently with these previous requirements. Thus, algorithms are usually of "low order", meaning that they do not store high derivatives of positions and velocities: this allows the time-step to be increased without invalidating the conservation of energy. However, numerical methods do not accurately follow the true trajectory for very long time. The unavoidable condition that nearby trajectories diverge exponentially from one another is exactly the chaos property discussed before. It results from the non-linearity of forcefield interactions. Nevertheless, there exists a procedure, the **Verlet algorithm**, which reduces the impact of this problem and makes it one of the most used in MD simulations.

1.2.2 The Verlet algorithm

There are many different versions of the original Verlet method[19]. The most known are the "leap-frog" [20] and the "velocity Verlet" [21] forms. Beyond the quoted fact about the low order in time, the Verlet algorithm is also quite easy to program, it can be reversed in time and it requires only one evaluation of the force per time-step. Here follows a pseudocode example of the algorithm, showing how it advances the coordinates `r` and momenta `p` over a **discrete time-step** `dt` via the knowledge of the forcefield `force(r)`.

```
do step = 1, nstep
 p = p + 0.5*dt*f
 r = r + dt*p/m
 f = force(r)
 p = p + 0.5*dt*f
enddo
```

Usually, in the MD framework, intramolecular bonds are not represented in terms of a potential energy function, mainly because of their very high vibration frequencies. Instead, the bonds are represented as **constraints**, forcing two bonded atoms to have a fixed distance. Constraints are introduced by the Lagrangian or Hamiltonian formalisms. Given an algebraic relation between two (atomic) coordinates, for instance a fixed bond length b between atoms 1 and 2, one may write a *constraint equation*, together with another equation for the first derivative (in time) of the constraint:

$$\chi(\vec{r}_1, \vec{r}_2) = (\vec{r}_1 - \vec{r}_2)^2 - b^2 = 0 \qquad \dot{\chi}(\vec{r}_1, \vec{r}_2) = 2 \cdot (\vec{v}_1 - \vec{v}_2) \cdot (\vec{r}_1 - \vec{r}_2) = 0 \tag{1.8}$$

For instance, in the Lagrangian context, the χ force acting on atom 1 and 2 enters in the form:

$$m_i \ddot{\vec{r}}_i = \vec{f}_i + \Lambda \vec{g}_i,$$

where Λ is a multiplier, $\vec{g}_1 = \frac{\delta \chi}{\delta \vec{r}_1} = -2 \cdot (\vec{r}_1 - \vec{r}_2)$ and $\vec{g}_2 = \frac{\delta \chi}{\delta \vec{r}_2} = 2 \cdot (\vec{r}_1 - \vec{r}_2)$. An exact expression of Λ can be recovered from the above equations. In case a number

[19] Verlet *Computer experiments on classical fluids.* Phys. Rev., 165, 201-214 (1968).

[20] Hockney et al. *Computer simulations using particles.* Adam Hilger, Bristol (1988).

[21] Swope et al. *A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: application to small water clusters* J. Chem. Phys., 76, 637-649 (1982).

M of constraints is adopted, a system of M equations is imposed. Since the equations of motion are solved approximately (dealing with discrete time-steps), the constraints are expected to be more and more violated as the simulation goes on. So an exact solution of such a system is not necessary. An interesting idea then is to determine the constraint forces in such a way that they are satisfied exactly just at the *end* of every time-step. When such a scheme is implemented within the original Verlet algorithm, it is called the SHAKE algorithm. The pseudocode becomes a bit more complicated, but still easy to be implemented:

```
do step = 1, nstep
 p = p + (dt/2)*f
 r = r + dt*p/m
 lambda_g = shake(r)
 p = p + lambda_g
 r = r + dt*lambda_g/m
 f = force(r)
 p = p + (dt/2)*f
 mu_g = rattle(r,p)
 p = p + mu_g
enddo
```

The `shake` routine calculates the constraint forces $\Lambda\vec{g}_i$ ensuring the first part of Eq. 1.8 is satisfied. The `rattle` routine calculates a new set of constraint forces $\Lambda\vec{g}_i$ ensuring the second part of Eq. 1.8 (time-derivative of the constraints forces) is satisfied at $\vec{r}_i$ at the end-of-step positions. In case of M constraints, these calculations are performed iteratively, in order to satisfy each constraint (in turn) until convergence. A simulation of a system with fixed bond lengths is obviously not equivalent to another with, say, harmonic constraints. The difference results in the distribution of the other (Lagrangian) coordinates. Calculating the configurational distribution function by integration over the momenta, the difference arises because in the former case a set of momenta is set to zero (and, thus, not integrated) while in the latter an integration is performed (leading to an extra term which depends on the particle coordinates). This issue is usually referred to as the *metric tensor problem*.

1.2.3 Non-bonded list

A non-bonded contribution to U^{tot} requires naturally a lot of pairwise calculations. To prevent all these calculations, it is possible to make the assumption of **short range interactions**. Consider two atoms i and j. The assumption of short-range interaction potentials means that $v(r_{ij}) = 0$ if $r_{ij} > r_{cutoff}$. In this case, the program skips the force evaluation. Nevertheless, the time to examine all pairs is still proportional to $\frac{1}{2}N(N-1)$ for N atoms. Considerable CPU time can be saved using a *list of nearby pairs*, a method suggested by Verlet himself (see Fig. 1.1). Each atom i is surrounded by a potential cutoff sphere of radius r_{cutoff}. Another sphere, of radius r_{list}, is then defined in such a way that at the first step an atom list is filled, containing all the j atoms whose distance from the i-th is less than r_{list}. In the next time-steps, only the i-j pairs appearing in this list are calculated by the `force` routine. The list must be refilled from time to time. The most important thing is to refill it before any unlisted pairs have crossed the region between r_{cutoff} and r_{list}. This can be done automatically if the distances covered by all the atoms (from the latest update) is recorded. The setting up of r_{list} must adapt to a compromise. The large lists need to be refilled less frequently, but save less CPU time than small lists.

1.2.4 Thermostats

In general there exist two approaches to perform MD simulations at constant temperature. The NVT- (micro canonical) *ensemble* is defined as a thermodynamics system in which the number of particles N, the volume V, and the temperature T are fixed. The temperature is defined by the *ensemble average of kinetic energies of all the N particles* and thus it is not possible to fix T exactly at each time-step. Therefore, a number of different **thermostats** – like Berendsen, Langevin and Nosè-Hoover – have been developed. A Berendsen thermostat[22] is a "proportional" type of thermostat: it corrects deviations of T from the set point T_0 (the "external bath") by multiply-

[22] Berendsen et al. *Molecular-Dynamics with Coupling to an External Bath.* J. Chem. Phy., 81, 8, 3684-3690 (1984).

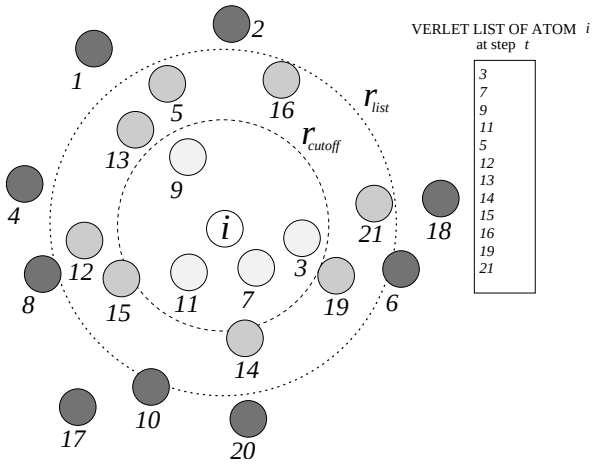


Figure 1.1: The Verlet list (of a fixed atom i in a selected time-step t): the list must be reconstructed before particles originally outside the list range (darkest spheres out of the dotted r_{list} circle) have entered the potential cutoff sphere (dashed r_{cutoff} circle). The force routine at step $t+1$ will calculate the r_{ij} distance only the for those atoms appearing in the list, in this case 12 atoms instead of 21: not too bad.

ing the atoms velocities by a certain factor, in order to control the value of T. The *Nosè-Hoover* thermostat introduces a *thermal reservoir variable* within the equations of motion[23], $\dot{\vec{p}}_i = f_i - \zeta \vec{p}_i$. It can be shown that if the system is too hot, then the "friction coefficient" ζ tends to increase; but if ζ is positive, the system cools down. Conversely, if the system is too cold ζ becomes negative and tends to heat up the system. Although sometimes this method leads to non-ergodic behaviours, this issue can be usually circumvented[24]. The Langevin thermostat relies on the Langevin equation of motion, rather than on the Newton's one[25]. In this framework, a "frictional force" is added to the conservative force (proportional to the velocity), adjusting the kinetic energy of the particle in such a way that the temperature becomes consistent with the set temperature (similarly to Nosè-Hoover thermostat). The *Andersen thermostat*, finally, adds a stochastic term to the temperature by simulating random collisions of the molecule atoms with a notional heat bath at the selected temperature: in practice,

[23] Nosè *A molecular dynamics method for simulations in the canonical ensemble.* Molec. Phys., 52, 255-268 (1984); Hoover *Canonical dynamics - equilibrium phase-space distributions.* Phys. Rev. A, 31, 1695-1697 (1985).

[24] Martyna et al. *Nose-Hoover chains: the canonical ensemble via continuous dynamics.* J. Chem. Phys., 97, 2635-2643 (1992).

[25] Adelman et al. *Generalised Langevin Equation Approach for Atom-Solid-Surface Scattering - General Formulation for Classical Scattering Off Harmonic Solids.* J. Chem. Phy. 64, 6, 2375-2388 (1976).

the velocity of a random particle is randomly reassigned from a Maxwell-Boltzmann distribution (at the selected temperature) to each component of the particle's velocity and added to the molecule at study[26].

1.2.5 Convergence of MD simulations

MD simulations make a finite sized molecular structure evolving in time, in a step-by-step way, and they have limitations in time and size scales. Time scale allowed by modern computer technology are typically of few nanoseconds to microseconds. Therefore, it is of crucial importance to assess whether or not a simulation has reached **equilibrium**, in order for average values calculated from it meaningful. Simulation averages need to be subjected to **statistical analysis** to estimate errors. How can we test that MD simulation have run long enough so that the results are well determined? From the experimental point of view the answer lies usually in repeating the experiment: reproducibility is, actually, a necessary (but not sufficient!) condition for good measurements.

Unfortunately these tests have not been commonly adopted by researchers, for it is computationally expensive. Available computer time is consumed simply by running simulations as long as possible, despite the fact that it has been proven that a large number of relatively short simulations are able to sample more fruitfully the phase space than a single, long trajectory[27].

MD is not an efficient way to sample thermodynamic basins. Thus, many techniques have been developed in order to enhance convergence, e.g. **umbrella sampling**[28] and **steered dynamics**[29]. These methods are useful but they too have their limitations. On the one hand they need to set *a priori* the reaction coordinate. Then, the superposition of restraint potentials invalidates the analysis of kinetic phenomenon. Eventually, both of the methods rely on sampling of all degree of freedom (DF) orthogonal to the

[26] Andersen, J. Chem. Phys. 72, 2384 (1980).

[27] Caves et al. *Locally accessible conformations of proteins: Multiple molecular dynamics simulations of crambin.* Protein Science, 7, 649-666 (1998).

[28] Patey et al. *A Monte Carlo method for obtaining the interionic potential of mean force in ionic solution.* J Chem Phys, 63, 2334-2339 (1975).

[29] Schulten *Manipulating proteins by steered molecular dynamics.* J Mol Graph Mod, 16, 289 (1998).

reaction coordinate and some of them can relax very slowly, inhibiting convergence. **Replica exchange** (or parallel tempering) MD combines multiple simulations at different temperatures with periodic exchange of temperatures[30], in such a way that the low temperature sampling is improved[31]. This methodology has its own side-effects. First, the number of replicas needed for a good sampling increases dramatically with system size and then, a great part of the CPU time is devoted to sample artificially high temperatures.

Conversely, standard MD still has great advantages. First of all, the kinetics of the system are realistic, namely in the context of the micro canonical ensemble or canonical ensemble in the presence of a thermostat. Secondly, there is no need to impose any reaction coordinate or biasing function at the beginning of a simulation, which means that the simulation is not "artifacted". Moreover, unperturbed MD can be matched with experiments that deal with both thermodynamic and kinetic aspects, For instance, magnetisation transfer experiments (a NMR technique) are related to the time correlation function of H-H distances. Therefore, the use of large scale MD is still of vital importance, above all since advances in computer technology have enormously improved the possibility of running long-time-scale MD simulations[32].

[30] See for instance U. H. E. Hansmann, *Computational Biophysics to Systems Biology*, Proceedings of the NIC Workshop (2008).

[31] Sanbonmatsu et al. *Structure of met-encephalin in explicit aqueous solution using replica exchange molecular dynamics.* Proteins: Struc Func Gen, 46, 225-234, (2002).

[32] Grossfield et al. *Convergence of Molecular Dynamics Simulations of Membrane Proteins*, PROTEINS: Structure, Function, and Bioinformatics, 67, 31-40, (2007).

Chapter 2

Solvation and solvation models

Still an introductory chapter, devoted to the basic principles of solvation, both from the physical and the computational point of view.

A little inaccuracy saves a world of explanation.
C. E. Ayres

A **solvent** is the component of a solution that is present in the greatest amount. It can be also defined as *a liquid substance in which other substances are dissolved in such a way that they can be removed from the solvent without changing their nature*[1] Among all the solvents, water plays an important role, since biological processes are possible only in this substance: in particular, proteins (but, in general, all the biological macromolecules) can perform their complex function (transport, folding, binding, catalysis) in aqueous solutions. Water affects all the physical and chemical properties of the dissolved biological material (the **solute**), from electronic densities to molecular association and therefore it is of great importance to accurately calculate the effect of the solvent in MD simulations.

The **solvation process** is the one in which *a particle of the solute is transferred from a fixed position in the gas phase into a fixed position in solution at constant temperature*[2]. The most important quantity which describes the effects of the solvent is the solvation energy ΔG_{sol}, defined as the reversible work which has to be spent in order to

[1] Marcus, *The properties of solvents*, Wiley, Chichester (1998).
[2] Ben-Naim *Standard thermodynamics of transfer. Uses and misuses*, J. Phys. Chem., 82, 792 (1978).

transfer the solute from the gas phase into solution[3]. From a microscopic point of view, the solvation effect arises because of three main reasons: (1) the *intra*molecular interactions within the solute; (2) the molecular interactions between solute and solvent; (3) the reorganisation of the solvent as a consequence of the stereochemical properties of the solute. The first two effects can be divided into two energy contributions[4], the **electrostatic** (or polar) contribution ΔG_{el} and the short-range or **van der Waals** contribution ΔG_{vdW}. The third one is commonly referred to as the **cavitation** (or non-polar) contribution to the solvation energy ΔG_{cav}. This partitioning of the solvation energy will be throughly investigated later.

2.1 Electrostatics

Electrostatic strength dominates the intramolecular interactions because of its magnitude and its long-range nature. The distribution of electrons around a certain atom nucleus creates a field which interacts with the one around another nucleus and these distributions play an important role in the solvation process. ΔG_{el} accounts for the the work needed to *create* the gas-phase of the solute charge distribution (in-solution) plus the work necessary to *polarise* it. This charge distribution polarises the surrounding solvent molecules, which generate a *reaction field* on the solute itself (and this field obviously affects the self-energy of solute atoms). Moreover, the intramolecular Coulomb-interactions are **screened** by the solvent molecules. Finally, the presence of ions in solution (such as Na+ and Cl-, coming from the presence of salt in water, or H+, due to the conditions of acidity/basicity) has a strong influence on conformational changes and binding properties of biological macromolecules[5].

[3] Orozco et al. *Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems*, J. Chem. Rev., 100, 4187 (2000).

[4] Cramer et al. *Implicit solvation models: equilibria, structure, spectra, and dynamics*, Chem. Rev., 99, 2161, (1999).

[5] For instance, Jovin et al. *The Transition Between B-DNA and Z-DNA*, Annu. Rev. Phys. Chem., 38, 521, (1987).

2.2 Short-range interactions

Let us briefly discuss the ΔG_{vdW} contribution. Van der Waals forces – which come from an effective dipole-dipole interaction between the solute and the solvent molecules[6] – are usually *favourable* to solvation, because dispersion forces are stronger than repulsive ones in a solute cavity.

The cavitation or *hydrophobic effect* can be defined as the energetic cost for creating a cavity in the solvent (needed to fit a solute molecule). It consists, in general, of an entropic contribution which takes into account the entropy loss resulting from the necessary rearrangement of the solvent molecules around the (nonpolar) solutes. In particular for water, it is related to the lowering in the number of favourable hydrogen bonds that water molecules could make between them if the solute (which does not allow hydrogen bonds) were not present. The term ΔG_{cav} is therefore *unfavourable* to solvation.

2.3 Free energy of solvation

In order to make a MD simulation computationally affordable it is necessary to make some assumptions about the complexity of the solvation effects. The most important assumption is that the free energy of solvation ΔG_{sol} can be partitioned in the following way:

$$\Delta G_{sol} = \Delta G_{el} + \Delta G_{vdW} + \Delta G_{cav}. \tag{2.1}$$

These different terms contribute in different ways to the total energy according to the different configurations of solvent and solute molecules. For instance, for pure water the dominant effect is ΔG_{el}, whereas the short-range term is of less importance. For an apolar solvent, like oil, the cavitation term is smaller: there is not a big energy loss in displacing the oil molecules, indeed, since they do not strongly interact to one another. On the other hand, polar solutes in polar solvents will result in a large electrostatic term, while for nonpolar solutes in nonpolar solvents the cavitation term will dominate (because ΔG_{el} is, in fact, close to zero).

[6] Cohen-Tannoudhi et al. *Quantum Mechanics*, Wiley Interscience Publications (1977).

2.4 Solvation models

As described in the previous chapter, an accurate representation of the aqueous solvent environment is fundamental in order to mimic the behaviour of soluble biomolecules. The most rigorous way to treat solvation effects is to include *explicitly* the molecules of solvent in the simulation system, but this leads to a very high computational cost. The solvent molecules greatly increase the number of degrees of freedom and the number of interactions between the system molecules. Simulations of proteins of about 100 residues in explicit water cannot sample nowadays more than 1 μs of real time. This limitation precludes the study of long-time-scale processes (e.g. protein folding), large scale structural transitions, multimeric assembly processes like complex formation and protein aggregation, as well as the derivation of accurate thermodynamical quantities (exactly because of the convergence problem quoted in the previous chapter).

This computational drawback has motivated the development of *implicit solvent models*[7]: they consist of a theoretical framework which captures the mean influence of solvent molecules around the solute by a **potential of mean force** that depends only on the atom coordinates of the solute[8] (see Fig. 2.1). The main advantages of an implicit solvent model are:

- it considerably reduces the system size;
- it avoids the need to average over the extremely large number of solvent configurations;
- it reduces the viscosity of the solvent environment by eliminating the friction from the solvent molecules, thus accelerating molecular motions[9];
- it directly yields the effective energy (the sum of the solute potential energy *in vacuo* and the solvation free energy)[10].

[7] Feig et al. *Recent advances in the development and applications of implicit solvent models in biomolecule simulations*, Current Opinion in Structural Biology, 14, 217-224 (2004); Roux et al. *Implicit solvent models*, Biophys. Chem., 78, 1-20 (1999); Bashford et al. *Generalised Born models of macromolecular solvation effects*, Annu. Rev. Phys. Chem., 51, 129-152 (2000).

[8] Cramer et al. *Implicit solvation models: equilibria, structure, spectra, and dynamics*, Chem. Rev., 99, 2161, (1999).

[9] Zagrovic et al. *Solvent Viscosity Dependence of the Folding Rate of a Small Protein: Distributed Computing Study*, J. Comp. Chem., 24, 1432-1436 (2003).

[10] In contrast, explicit water simulations have to be post-processed, for example by finite-difference Poisson-Boltzmann

Disadvantages of such models lie in the difficulty of *parametrisation*, which actually has to be made via MD simulations themselves and can thus lead to inconsistencies with experimental quantities. Moreover, these models do not account for all those phenomena strictly related to the *granular nature* of water such as structural water molecules (encompass active site water) and desolvation barriers (related to the entropy of water hydrogen bonds). Anyway, implicit solvent models can be classified into three main families:

1. Surface area models (SASA)[11];
2. Gaussian solvent-exclusion models (GASE)[12];
3. Dielectric continuum electrostatics models, or Continuum Treatment of Solvation (CTS).

2.5 CTS models

One way to implement the idea of an overall potential reproducing the effects of water on a macromolecule is to consider the surrounding water as a continuous medium with an high dielectric constant ϵ_{sol} and the macromolecule as a continued medium with a low dielectric constant ϵ_{mol}. These models are called dielectric continuum electrostatics or simply continuum dielectric models. They can be classified into finite-difference Poisson-Boltzmann (PB)[13] and generalised Born (GB)[14] models.

Generally speaking (further details will be given in the next section), PB models are more accurate than GB models, but they require more computation time and they

calculations, to obtain the effective energy.

[11] Eisenberg et al. *Solvation energy in protein folding and binding*, Nature, 319, 199-203 (1986); Ooi et al. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides, PNAS, 84, 3086-3090 (1987); Fraternali et al., *An Efficient Mean Solvation Force Model for Use Molecular Dynamics Simulations of Proteins in Aqueous Solution*, J. Mol. Biol., 256, 939-948 (1996). Ferrara et al., *Evaluation of a fast implicit solvent model for molecular dynamics simulations*, Proteins, 46, 24-33 (2002)

[12] Stouten et al. *An effective solvation term based on atomic occupancies for use in protein simulations*, Mol. Simul., 10, 97-120 (1993); Lazaridis et al. *Effective energy function for proteins in solution*, Proteins, 35, 133-152 (1999).

[13] Warwicker et al. *Calculation of the electric potential in the active site cleft due to α-helix dipoles*, J. Mol. Biol., 157, 671-679 (1982).

[14] Constanciel et al., Theor. Chim. Acta, 65, 1 (1984); Still et al. *Semianalytical treatment of solvation for molecular mechanics and dynamics*, J. Am. Chem. Soc., 112, 6127-6129 (1990).

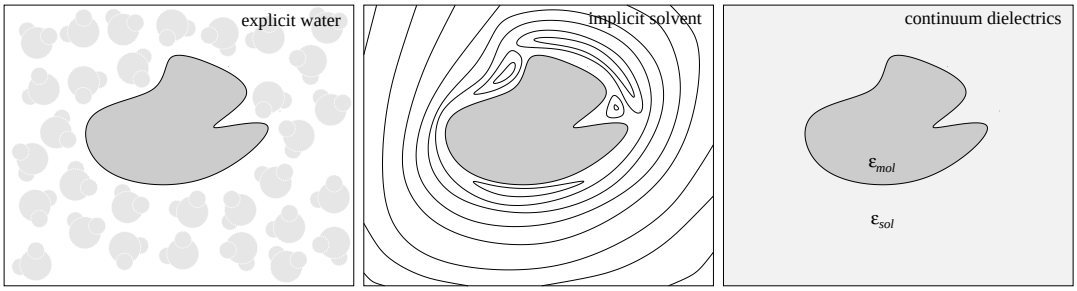


Figure 2.1: Progressive approximation of a macromolecule's aqueous environment according to explicit water, implicit water and continuum dielectrics implicit water models. In the first case (left), the solvation effect is accounted for by explicit calculation of water molecules interactions (light grey) surrounding the macromolecule (dark grey); in the second case (middle), the water surrounding is approximated by an overall potential (illustrative equipotential lines are shown); in the last case (right), the potential of mean force is obtained by the interaction of two continuum dielectrics: one (light grey) representing the aqueous environment (characterised by a dielectric constant ϵ_{sol}) and the other one (characterised by a dielectric constant ϵ_{mol}) representing the macromolecule (dark grey).

suffer of difficulties in the derivation of forces. The GB model is related to the PB model but contains several approximations that increase the calculation speed. There exist models that combine different approximations like GBSA[15], where the polar part is treated through GB formalism and the nonpolar part through a surface area term.

2.5.1 Electrostatics: PB equation and GB models

The water surrounding a macromolecule is replaced by a continuous and uniform medium with a high dielectric constant $\epsilon_{sol} = 80$ (in units of ϵ_0 here and henceforth); the macromolecule is replaced by a region with low dielectric constant ($\epsilon_{mol} = 1$). A spatial charge distribution $\rho(\vec{r})$ within the low dielectric medium emulates the partial charges q_i of molecule atoms. In this context there are two energetic contributions to take into account: the *screened interactions* (the interactions between the solute molecules to one another) and the *solvation energy* (or *self energy*, coming from the interactions between the solute-solvent interactions). The first term consists of the

[15]Schaefer et al. *A Comprehensive Analytical Treatment of Continuum Electrostatics*, J. Phys. Chem., 100, 1578-1599, (1996); Im et al., *Generalised Born model with a simple smoothing function*, J. Comput. Chem., 24, 1691-1702 (2003); Mongan et al. *Generalised Born Model with a Simple, Robust Molecular Volume Correction*, J. Chem. Theory Comput., 3, 156-169, (2007).

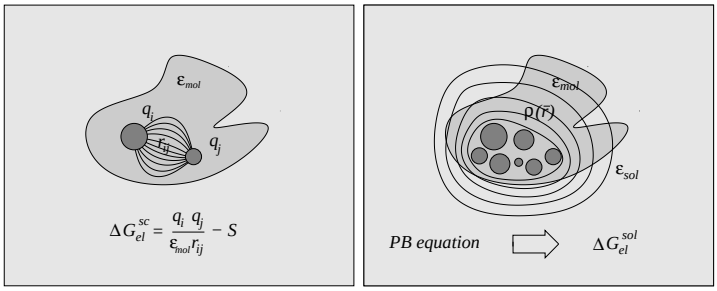


Figure 2.2: Screened interactions (left) and solvation energy contributions (right).

canonical Coulomb-interaction damped by a factor that takes into account the dielectric interposed between the charge pairs $\Delta G_{el}^{sc} = \sum_{i,j=1}^{N} \frac{q_i q_j}{\epsilon_m r_{ij}} - S$. The second one, ΔG_{el}^{sol}, consists of the interactions of partial charges with the induced solvent reaction field, calculated on the basis of the Poisson-Boltzmann equation (see Fig. 2.2).

2.5.2 The Poisson-Boltzmann equation

The dielectrics ϵ_{sol} and ϵ_{mol} are assumed to be *constant* and *real* in their regions, meaning that water and molecular matter are assumed to be homogeneous dielectrics and not subjected to any resonance effects, respectively. Further hypothesis about the dielectric's features concern *linearity* (see Fig. 2.3). In fact, the polarisation vectors $\vec{P}_{sol}$ and $\vec{P}_{mol}$ are both assumed *proportional* to an external electric field $\vec{E}$. Indeed, $\vec{P} = \epsilon_0(\epsilon_r - 1)\vec{E}$ and $\vec{D} = \epsilon_0\vec{E} + \vec{P}$, implies $\vec{D} = \epsilon_0\vec{E} + \epsilon_0(\epsilon_r - 1)\vec{E} = \epsilon\vec{E}$. Combining these results with the elementary relation $\vec{E} = -\vec{\nabla}\phi$, where ϕ is the electrostatic potential, it reads: $\vec{D} = -\epsilon\vec{\nabla}\phi$. By the second Maxwell equation $\vec{\nabla} \cdot \vec{D} = 4\pi\rho$, hence:

$$\vec{\nabla} \cdot (\epsilon\vec{\nabla}\phi) = -4\pi\rho, \tag{2.2}$$

also known as the *Poisson-Boltzmann equation*. This relation allows to define the electrostatic potential ϕ given the charge distribution ρ. With the assumption of constant dielectric this yields:

$$\nabla^2\phi(\vec{r}) = -\frac{4\pi}{\epsilon}\rho(\vec{r}). \tag{2.3}$$

This way the Poisson-Boltzmann equation becomes a *Poisson equation*, a very well studied second order partial differential equation[16]. It requires two initial conditions. One choice is to fix these conditions in such a way that, for $r \to \infty$, it will result in:

$$\begin{cases} \phi(\vec{r}) \sim & \alpha/r \\ \nabla\phi(\vec{r}) \sim & \beta/r^2 \end{cases}$$

where α and β are finite real numbers and $r = |\vec{r}|$. The electrostatic contribution to the solvation free energy can be written as a function of the potential ϕ:

$$\Delta G_{el}^{sol} = \frac{1}{2} \int \rho(\vec{r})\phi(\vec{r})d\vec{r} \tag{2.4}$$

The continuum dielectrics approximation is satisfactory if the dipoles of the solvent are far smaller than the ones of the solute. This is obviously not true in the case of macromolecules, whose characteristic dipoles are similar in magnitude to the ones of water molecules (see Fig. 2.3)[17]. Two crucial points are the implementation a *polarisable forcefield* and a more refined treatment of the nonpolar contribution to the solvation energy (which will be discussed later).

One problem to be solved in a PB model, is how to define the boundary between molecule and solvent. This leads to the definition of the "van der Waals envelope" and the "solvent accessible surface", which will be briefly discussed later. There is naturally a *discontinuity* between these two regions: but since a **finite differences** method is used to solve this differential equation it is possible to interpolate values of the dielectric constants across the boundary.

2.5.3 Finite differences method for solving the Poisson equation

As we have seen, given the dielectric function $\epsilon(\vec{r})$, the electrostatic potential $\phi(\vec{r})$ of a charge distribution $\rho(\vec{r})$ is uniquely defined by Eq. 2.2, provided linearity of the

[16] Evans *Partial Differential Equations*, American Mathematical Society, Providence, (1998); Polyanin *Handbook of Linear Partial Differential Equations for Engineers and Scientists*, Chapman & Hall/CRC Press, Boca Raton (2002).

[17] Grycuk *Deficiency of the Coulombic-field approximation in the generalised Born model: an improved formula for Born radii evaluation*, J. Chem. Phys., 119, 9, 4817-4826 (2003); Lwin et al. *Is Poisson-Boltzmann theory sufficient for protein folding simulations?*, J. Chem. Phys., 124, 034902-1-03492-6 (2006); Wang et al. *Poisson Boltzmann solvents in molecular dynamics simulations*, Comm. Comp. Phys., 3, 5, 1010-1031 (2008).

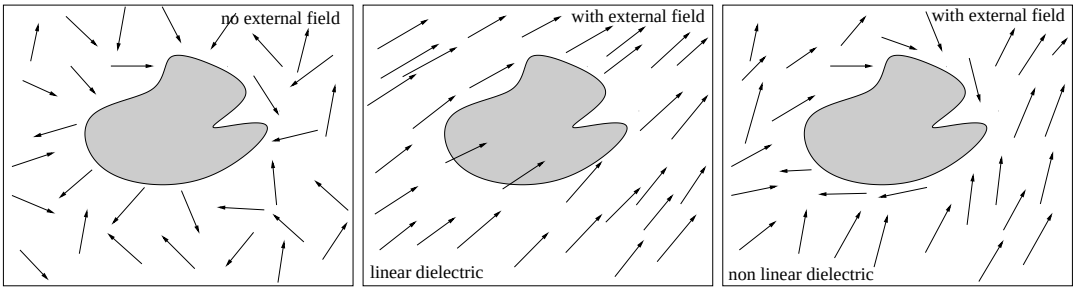


Figure 2.3: Different response to a external electric field of a dipoles system (water). Left panel shows the water dipoles moving around a macromolecule in absence of any external field. (Center) The external field (which is directed from the bottom left to the top right of the box) response of a dipole system which can be considered a linear dielectric. The dipoles are displaced in such a way that the polarisation vector (the average vector of the single dipoles in the selected volume, e.g. the box) is *everywhere* parallel to the external field: $\vec{P} \simeq \epsilon_0 \chi \vec{E}$. (Right) Response of a non linear dielectric. In proximity of the macromolecule boundary, because of the similar strength in the dipoles of the macromolecule and local water, the water dipoles do not rearrange in such a way that the polarisation vector is parallel to the external field: thus, $\vec{P} \neq \epsilon_0 \chi \vec{E}$.

polarisation vectors. Now, in order to obtain the solvation energy ΔG_{el}^{sol} this equation must be *integrated*. As often occurs with integrals, no analytical solutions are available. It can be solved only numerically. The most popular numerical method to solve such an equation is the the *finite differences* (Poisson) method (fdP)[18]. Here, we do not describe the details of this technique[19]. The main point is that this is a very inefficient method. In fact, in the context of a MD simulation, one has to remember that the fdP routine has to be called *at each time-step*. Furthermore, it is difficult to derive forces due to the discrete nature of the solvation process. However, although fdP methods are very time demanding, it is actually the most accurate implicit solvent model, despite all the approximations. It represents a benchmark for all other implicit solvent models. Moreover, the method is still more efficient than explicit water.

[18] Morton et al. *Numerical Solution of Partial Differential Equations, An Introduction*, Cambridge University Press, 2005; Rübenkönig *The Finite Difference Method - An introduction*, Albert Ludwigs University of Freiburg (2006).

[19] See for instance Zhou et al. *Finite-Difference Solution of the Poisson-Boltzmann Equation*, J. Comp. Chem., 11, 11, 1344-1351 (1996).

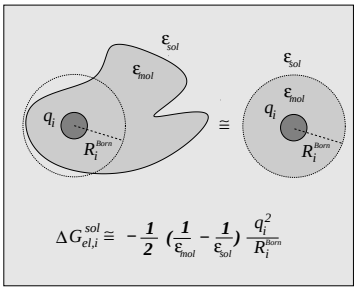


Figure 2.4: Definition of the *effective Born radius* of atom i with (partial) charge q_i, absorbed in a dielectric medium with dielectric constant ϵ_{mol} and surrounded in turn by a dielectric medium with dielectric constant ϵ_{sol}. If the main energetic contribution of such a system comes from the dielectric contained in a sphere of radius R_i^{Born} around atom i (left), then it is possible to approximate the system with a completely spherical system (right): in this case, the computation of the electrostatic solvation energy $\Delta G_{el,i}^{sol}$ becomes a quite trivial task, being simply the superposition of the effects of two spherical capacitors.

2.5.4 GB models

GB models became popular because of their balance between their relative accuracy to approximate the fdP solvation energies and their lower computational cost. These models treat the problem of continuum dielectric with the definition of an *effective Born radius*, R_i^{Born}, for each atom (see Fig. 2.4). Suppose all the charges are switched off, apart the one related to the atom i, with charge q_i within the molecule dielectric. Let $\Delta G_{el,i}^{sol}$ be its atomic self energy (that can calculated by fdP). $\Delta G_{el,i}^{sol}$ can be approximately recovered by replacing such a system by a *sphere* with charge q_i and radius

$$R_i^{Born} = -\frac{1}{2}\underbrace{\left(\frac{1}{\epsilon_{mol}} - \frac{1}{\epsilon_{sol}}\right)}_{=\tau=\frac{1}{1}-\frac{1}{80}=\frac{79}{80}\simeq 1}\frac{q_i^2}{\Delta G_{el,i}^{sol}}. \tag{2.5}$$

Such an approximation is supported by the fact that electrostatic interactions, once we have fixed a uncertainty threshold of accuracy, can be neglected beyond a certain distance. The introduction of such a R_i^{Born} radius is useful since it can be shown that the *total* electrostatic solvation energy can be written as a function of these radii in

the following manner:

$$\Delta G_{el}^{sol} \simeq -\frac{\tau}{2}\sum_{i,j=1}^{N}\frac{q_i q_j}{\sqrt{r_{ij}^2 + R_i^{Born}R_j^{Born}e^{-\frac{1}{4}\frac{r_{ij}^2}{R_i^{Born}R_j^{Born}}}}} \tag{2.6}$$

The key step is to calculate the quantity R_i^{Born}. The most accurate way would be to derive it from fdP calculations[20], but his would bring no computational advantages. Popular GB models make use of spatial integrals.

2.5.5 Nonpolar interactions: solvent accessible surface models

The study of the protein folding problem[21] and hydrophobicity[22] motivated a deeper understanding of surface-related effects and the development of several "protein surface" definitions. The current model of a folded protein[23] assumes hydrophobic sidechains to be preferentially buried away with respect to the external aqueous solvent. Quantitative analysis of this hydrophobic burial effect[24] led to the concept of *solvent accessible surface* (SASA)[25]. The relevance of such studies comes from the fact that (in a first approximation), the hydrophobic effect is (experimentally) proportional to the SASA[26]. Since the precise definition of SASA is not trivial, we consider some details about it. It is important to remember that the surface and the volume of a molecule are, actually, *abstract* objects. A molecule is a dynamical system in which the particles are referred to as time-dependent probability distributions of presence of mass (orbitals), held together by electromagnetic force and, for these very reasons, they do not have any boundaries.

[20] Scarsi et al. *Comparison of a GB solvation model with explicit solvent simulations: potentials of mean force and conformational preferences of alanine dipeptide and 1,2-dichloroethane*, J. Phys. Chem. B, 102, 3637-3641 (1998). David et al., *Comparison of generalised Born and Poisson models: energetics and dynamics of HIV Protease*, J. Com. Chem, 21, 295-309 (2000).

[21] Anfinsen *Principles that Govern the Folding of Chains*, Science, 181, 223-230 (1973).

[22] Tanford *The Hydrophobic Effect*, Wiley Interscience, New York (1980).

[23] Kauzmann *Some factors in the interpretation of protein denaturation*, Adv. Protein Chem., 14, 1-63, (1959).

[24] Chothia *Hydrophobic bonding and accessible surface area in proteins*, Nature, 248, 338-339 (1974).

[25] Lee et al. *The interpretation of protein structures: Estimation of static accessibility*, J. Mol. Biol., 55, 379-400 (1971).

[26] Eisenberg et al., Nature, 319, 199 (1986); Ooi, et al., Proc. Natl. Acad. Sci. U.S.A. (1986); Hermann, Phys. Chem., 76, 2754 (1972); Amidon et al., Phys. Chem., 72, 2239 (1975); Floris, J. J. Comput. Chem., 10, 616 (1989).

Van der Waals envelope

Two not covalently bound atoms can not approach each other closer than a certain distance due to electron shell repulsion. The maximal proximity depends on the type of the involved atoms. This fact can be described by assigning a *van der Waals radius* r_W to each atom type, in such a way that the sum of these quantities (for a given atom pair, say i and j), is equal to their closest distance d_{ij}, $r_W^i + r_W^j \leq d_{ij}$. We refer to the *union* of these (spherical) atomic surfaces as the *van der Waals envelope*.

Solvent accessible surface

Sometimes the van der Waals envelope can be misleading. Macromolecules frequently display small gaps and clefts, which are actually too small to accommodate a water molecule. Thus, the van der Waals surface of these pockets can not contact a solvent or a ligand: therefore it is not an "accessible surface" for the solvent. To "smooth" the roughness of the van der Waals envelope the concept of a "contact surface" and a "solvent accessible surface" were then introduced[27]. These surfaces are obtained by rolling a spherical probe of a diameter equal to the size of a water molecule on the van der Walls envelope. In particular:

- The area where the probe touches the van der Waals envelope is called the *contact surface*;
- The collection of the probe's central points is called the *solvent accessible surface*;
- The patches over the clefts traced by the surface of the probe are called *re-entrant surfaces*;
- Contact surface + re-entrant surface is then referred to as the *molecular surface*.

The implementation of SASA theory in MD simulations was made possible by a very efficient computer algorithm for deriving these surfaces[28]. The most fruitful applications of this theoretical work is the linear treatment of the nonpolar contribution to

[27] Richards *Areas, volumes, packing and protein structure*, Annu. Rev. Biophys. Bioeng., 6, 151-176 (1977).

[28] Connolly *Analytical molecular surface calculation*, J. Appl. Crystallogr., 16, 548-558 (1983); Connolly *Solvent-accessible surfaces of proteins and nucleic acids*, Science, 221, 709-713 (1983).

the free solvation energy of an atom i:

$$\Delta G_{np,i}^{sol,sasa} = \Delta G_{cav,i}^{sol} + \Delta_{vdW,i}^{sol} = \gamma \cdot S_i, \tag{2.7}$$

where γ is the coefficient (usually referred as to the *surface tension*, in units of [kcal/(mol $\cdot$ Å^2)]) by which, when multiplied by the contribution of atom i to the total SASA of a macromolecule S_i, ones obtains the corresponding nonpolar contribution to the free solvation energy $\Delta G_{np,i}^{sol}$.

2.5.6 Beyond SASA: the Tolman theory

Let Σ be the surface of a macromolecule and S its related SASA. This (ideal) surface has a peculiar *curvature* in each point $\vec{r}$ at the interface with water $\sigma(\vec{r})$, which can be defined by means of the more intuitive radius of curvature $\rho(\vec{r}) = 1/\sigma(\vec{r})$. The Tolman theory of surface tension[29], a second order approximation of SASA theory, states that the surface tension γ can be written as:

$$\frac{\gamma[\rho(\vec{r})]}{\gamma[\rho(\vec{r}) \to \infty]} = \frac{1}{1 \pm a/\rho(\vec{r})} = \frac{1}{1 + a \cdot \sigma(\vec{r})},$$

where a is the radius of a water molecule ($a = 1.4$ Å for our purposes) and the sign depends on the concavity of the curvature (here, σ is negative for *convex* cavities, following Tsodikov[30]), leading to a more refined expression of Eq. 6.1, provided the substitution $\gamma \simeq \gamma[\rho(\vec{r}) \to \infty]$:

$$\Delta G_{np,i}^{sol,tol} = \frac{\gamma \cdot S_i}{1 \pm a/\rho(\vec{r}_i)} = \frac{\gamma \cdot S_i}{1 + a \cdot \sigma(\vec{r}_i)}. \tag{2.8}$$

In order to evaluate the effect of the Tolman correction, an estimation of the difference between $\Delta G_{np,i}^{sol,sasa}$ and $\Delta G_{np,i}^{sol,tol}$ is useful[31]. By means of the `surface_racer_5.0`

[29] Tolman, J. Chem. Phys. 17, 333 (1949); Salvino, Phys. Rev. B, 34, 6351-6366 (1986); Sharp et al. *Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects*, Science, 252, 106-109 (1991); Sharp et al. *Extracting hydrophobic free energies form experimental data: relationship to protein folding and theoretical models*, Biochemistry, 30, 9686-9697 (1991); Su et al. *A continuum approach to microscopic surface tension for the n-alkanes*, Ind. Eng. Chem. REs., 35, 3399-3402, (1996); Rashke et al. *Quantification of the hydrophobic interaction by simulations of the aggregation of small hydrophobic solutes in water*, PNAS, 98, 11, 5965-5969 (2001); Markin et al. *Quantitative theory of surface tension and surface potential of aqueous solutions of electrolytes*, J. Phys. Chem B., 106, 11810-11817 (2002);

[30] Tsodikov et al. *A novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature*, J. Comput. Chem., 23, 600-609 (2002).

[31] A hypothetical software implementation of such a correction into an existing model is easier this way, rather than rewriting the entire nonpolar module, from the programmer's point of view.

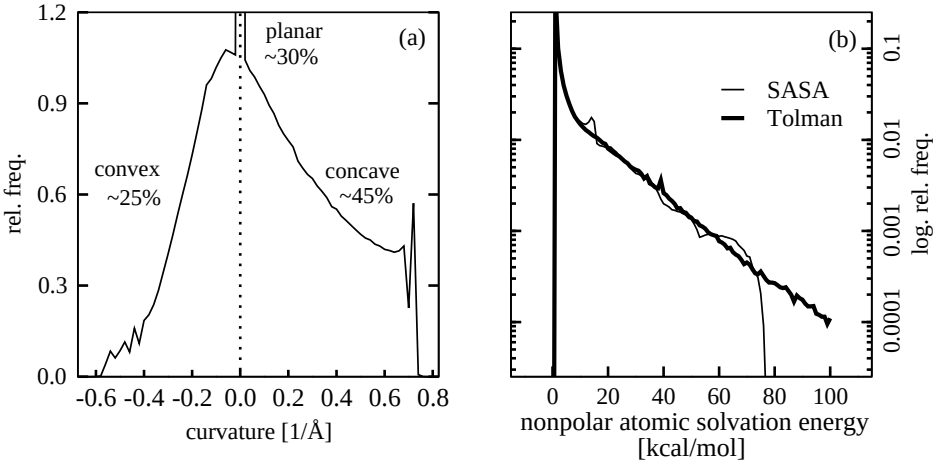


Figure 2.5: (a) Distribution of (local) curvatures σ_i in the testcase obtained with the surface_racer_5.0 program: it resulted in a significant skewness towards concave curvatures, although a large number of the atoms gives a null contribution (their Σ_i is almost planar). (b) Comparison between $\Delta G_{np,i}^{sol,sasa}$ and $\Delta G_{np,i}^{sol,tol}$ (written with $\gamma = 1$ [kcal/(mol $\cdot$ Å^2)] for sake of simplicity) on the basis of the S_i computed by CHARMM over all the protein testcases in relation with the (local) curvatures σ_i seen in plot (a): SASA tends to globally *underestimate* the solvation energy with respect to Tolman model.

software tool[32], we computed the local surface curvature σ_i (i.e. the curvature of Σ_i, the contribution to Σ coming from the atom i) correspond to each atomic position ($\rho(\vec{r}_i) \rightarrow i$) of the following peptide-protein testcase: 1crn 1cus 1dvd 1enh 1f8a 1fmk 1hdn 1hel 1inc 1kvd 1l2y 1lz1 1pgb 1pht 1shg 1ubq 1ycq 1ycr 2a3d 2ci2 2ins 2ptl 3app 3pte 5hvp anki bet1 gsgs hlx1 ins2 prph (each in 101 different conformations); these are very different structures with widely different solvation properties[33]. Results of such a study are summarised in Fig. 6.4. It shows that $\Delta G_{np,i}^{sol,sasa}$ is more *favourable* for convex cavities than $\Delta G_{np,i}^{sol,sasa}$, meaning that the Tolman-corrected model tends to *unfavour* solvation of convex cavities (the ones pointing *out* of the macromolecule) more than the popular SASA approach; viceversa, Tolman's model enhances the solvation of concave surfaces (e.g. re-entrant surfaces, and in general all the pockets whose curvature is oriented towards the *interior* of the protein) with respect to SASA: which makes sense, since Tolman's theory, by means of the knowledge of local $\sigma(\vec{r})$, discriminates between

[32] Tsodikov *ibidem*.

[33] Coleman et al. *An intuitive approach to measuring proteins surface curvature*, PROTEINS, Stru. func. bio., 61, 1068-1074 (2005).

regions of Σ which allow a spherical water probe (of radius a) to fit in and regions which do not. This is in agreement with Eq. 6.2.

The effect is 4–5 times more relevant (from a free energy point of view) for convex cavities rather than for concave (negative values of $\Delta G_{np,i}^{sol,sasa}$-$\Delta G_{np,i}^{sol,tol}$ span between -200 kcal/mol to 0, with more than 30% of the value between -50 and 0 kcal/mol, while positive values of the same difference range between 0 and 35 kcal/mol, with more than 50% between 0 and 10 kcal/mol).

Chapter 3

The FACTS model

This is a detailed overview of the FACTS implicit solvent model in his original form (i.e. without any correction). The peculiar geometric methods exploited by FACTS to derive the free energy of solvation are discussed.

Everything we call real is made of things that cannot be regarded as real.
N. Bohr

The name of the implicit solvent model (**F**ast **A**nalytical **C**ontinuum **T**reatment of **S**olvation), the subject of this work, immediately suggests two ideas. The first is that we are dealing with a -**CTS** model (in particular, FACTS is a GB model). Second, the main purpose of such a model is to *speed up* the MD simulations. Such an improvement of the dynamics is attained by reducing the time needed to calculate Born radii. FACTS actually exploits a **fitted function** to calculate them (while they are usually recovered via the very time demanding fdP method or some sort of volume integrals) without loosing precision. An accurate evaluation of these quantities is crucial for a good GB models[1]. The present chapter is devoted to the study of such a function and to the geometric quantities at the basis of the fitting procedure.

[1] Onufriev et al., J. Comp. Chem., 23, 1297-1304, (2002).

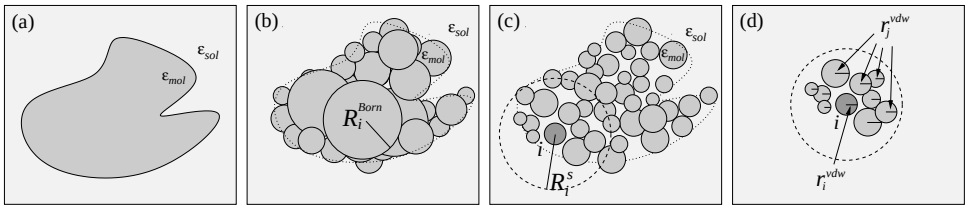


Figure 3.1: (a) In a continuum dielectric model the border between water and molecule dielectrics is essentially the molecular surface; in the GB models (b), a group of spherical atoms (where the sphere can be thought of about the R_i^{Born} radius) constitutes the macromolecular dielectric with dielectric constant ϵ_{mol}. The macromolecule as a geometric group of atoms (c): the atom i and its neighbouring atoms belonging to a surrounding sphere R^s. (d) The encumbrance of each atom in the R^s sphere is assumed to be related to its van der Waals radius r_j^{vdW}. The summation of the volumes of these spheres, including the central one with r_i^{vdW}, gives an idea of the amount of water flushed out off of the R^s sphere.

3.1 The point of view

In the context of a continuum dielectric model the geometric properties of the region surrounding a given atom in a macromolecule are almost neglected, since the focus is on the characteristics of the *boundary* between the two dielectrics. In the SASA implicit solvent model[2], for instance, all the solvation features of a biological molecule are defined by the knowledge of the solvent accessible surface. GB models face the problem of analytically computing the solvation energy of a macromolecule/solvent dielectric environment as follows. The solute is treated as a group of spherical objects of radius R_i^{Born}, each made up of a (solute) dielectrics characterised by dielectric constant of ϵ_{mol}, immersed in another (solvent) dielectric characterised by a dielectric constant of ϵ_{sol}, interacting with one another in such a way that the total energy is the one of Eq. 2.6. Thus GB models completely neglect the *boundary* between the dielectrics, see Fig. 3.1 (a,b). The most time demanding step of a GB model is the evaluation of the *atomic* electrostatic solvation energies, needed to recover the Born radii (Eq. 2.5), which are in turn used in Eq. 2.6 to calculate the total solvation energy. At this point it is important to point out a characteristic of all GB models. In the neighbourhood of each atom the (solute) uniform dielectric fills the space *completely*. The consequences

[2]Ferrara et al. *Evaluation of a fast implicit solvent model for molecular dynamics simulations*, Proteins, 46, 24-33 (2002).

of this property are relevant in the following, and henceforth we will refer to it as the "horror vacui hypothesis".

A more geometric, rather than physical, representation of the macromolecule immersed in an implicit water environment (like the one shown in Fig. 3.1 (c)) allows us to stress the relevance of the *spatial encumbrance* of atoms. In such a simple hard-spheres model (where the radii of the spheres representing atoms are their van der Waals radii), let us fix the attention on a given atom i and on its neighbours contained in a sphere of radius R^s with the center set on the atom i: it is clear that:

- the more the R^s sphere is **filled up** by other atoms, the more unlikely it is to find water in it (i.e. the atomic solvation energy atom i should decrease);
- the more **symmetric** the distribution of atoms around i is, given the same number and type of atoms within R^s (but in different dispositions), the more atom i becomes unaccessible to water (i.e. again, the atomic solvation energy atom i should decrease).

These two ideas will be studied in depth in the next sections. Now, it is important to point out that the FACTS model will exploit these geometrical observations to infer the Born radii (at each timestep) on the basis of a *statistical sample*, rather than by calculating atomic solvation energies by means of the (expensive) fdP procedure. Since calculating the value of such a given function is less time demanding than numerically solving the Poisson equation, FACTS will result faster than all those GB models – hence, the "F" in the FACTS acronym.

3.2 Excluded volume and neighbourhood symmetry

3.2.1 The volume measure A

The idea of water displacement quoted above is rather rough at this stage. Here, a more refined definition of such a measure connecting geometric information and solvation energy is studied. Fig. 3.1 (d) suggests that it is possible to define a *quantitative measure* of the "water displacement" around the fixed atom i quoted above: we will call this measure A, in units of Å^3. As a first approximation, A_i can simply be set

equal to the *summation of the van der Waals volumes* $(= 4/3\pi \cdot (r_j^{vdW})^3$ where r_j^{vdW} is the van der Waals radius of atom j and $j = 1 \dots N$ being the number of neighbouring atoms inside the sphere of radius R^s centered on atom i) *of the surrounding atoms in the sphere of radius* R^s *around atom* i. The radius R^s must be chosen carefully: if it is too small, the "water displacement" measure is obviously too inaccurate, but if it is too large, lots of the inner atoms do not significantly affect A_i (in the Introduction we understood how important it is to reduce the number of atoms to be computed in order to speed up the dynamics).

R^s should then be chosen as small as possible for efficiency reason, but should be large enough to derive meaningful/accurate "water displacement". A different R^s radius is given for each vdW radius of the central atom i, since its encumbrance affects the way the surrounding atoms can dispose around it. Actually this kind of ideas recalls the *sphere packing problem*, i.e. calculate the number of identical size sphere in a crate[3], with two additional conditions. The sphere radii are different and the box is spherical, but here we will not deal with such a problem.

In the original version of the model there are 7 different vdW radii and thus 7 different R^s radii (for the CHARMM parameter 19 united atom force field), meaning that FACTS recognises 7 different atom types (see Tab. 3.1 for details). The definition of the measure A of a certain atom i can then be given as $A_i = \sum_j^N V_j$, where N is the number of atoms within the $R^s_{(i)}$ sphere assigned by Tab. 3.1 and V_j the vdW volume of each atom. Such a measure, on the other hand, is clearly *not analytical*, meaning that it is impossible to define its derivatives (first and second) needed to perform MD calculations, because of the discrete nature of its definition. The issue can be overcome by means of a smoothing function Θ:

$$\Theta(r_{ij}) = \Theta_{ij} = \begin{cases} \left[1 - \left(\frac{r_{ij}}{R^s}\right)^2\right]^2 & r_{ij} \leq R^s \\ 0 & r_{ij} > R^s \end{cases} \tag{3.1}$$

Beside, the Θ function weights the contribution to solvation according to the distance of the j atoms from the center (this makes sense, since the further an atom is from i,

[3]Sloane *The sphere packing problem*, Docu. Math., 3, 387-396 (1998); Hales *The sphere packing problem*, Phys. Rev. Lett., 47, 1121-1124 (1981) and J. Comp. App. Math., 44, 41-76 (1992); Manoharan, et al. *Dense Packing and Symmetry in Small Clusters of Microspheres*, Science, 301, 483 (2003).

TYPE	r^{vdW} [Å]	R^s [Å]
H*	1.0000	7.39032
N* O*	1.6000	8.46133
S	1.8900	9.17618
C	2.1000	9.58532
CH3E	2.1650	10.00000
CH2E	2.2350	9.39675
CH1E	2.2365	10.00000

Table 3.1: FACTS atom types: according to their van der Waals radius, the FACTS R^s sphere is assigned. R^s has been parametrised in such a way that ΔG_i^{el} does not change significantly under the effects of a change of atom disposition outside a sphere of such a radius (these R^s value are related to $\epsilon_m = 2$ and $\epsilon_s = 78.5$). The parametrisation was indeed carried out comparing the solvation energy of a R^s sphere completely surrounded by solute and the one of a R^s sphere completely surrounded by solvent, looking for the minimum R^s for which the difference between these two energies were negligible (and this for *any* atom conformation *within* the sphere). The symbol "*" after H, N and O indicates that all the CHARMM param19 atom types whose name begins with the respective letter are included in that FACTS atom type.

the less it affects its solvation properties). Eventually, the measure

$$A_i = \sum_{j \in R^s}^{N} V_j \cdot \Theta_{ij} \tag{3.2}$$

becomes analytical (hence the "A" in the FACTS acronym). The same procedure will now be applied to the symmetry measure.

3.2.2 The symmetry measure B

As noted before, the probability of an atom to be surrounded by water is highest when no solute atoms are present in the R^s sphere, but we also pointed out that this probability is higher if the distribution of the surrounding atoms is **asymmetric** as well. The FACTS measure of symmetry B_i consists of a volume-weighted, normalised sum of the versors $\hat{x}_{ij}$ pointing from the central atom i to the N neighbouring atoms

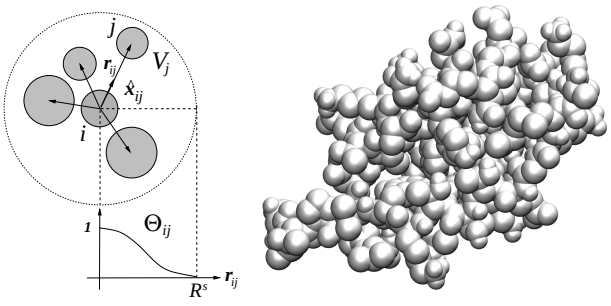


Figure 3.2: Schematic view and "real" aspect of the FACTS ingredients. In the left scheme, the R^s sphere and some atoms j surrounding the central one i are shown. The smoothing function Θ_{ij} dampens the contribution to A_i and B_i. The $\hat{x}_{ij}$ versor connects the i and j centres. V_j is the van der Waals volume of atom j. (Right) A more realistic view of the atom disposition (protein G, 1igd code): spheres radii are the van der Waals radii of each atom: remarkably, a large **overlap** between spheres is present.

j, weighted by the Θ_{ij} smoothing function.

$$B_i = \left| \frac{\sum_{j=1, j\neq i}^{N} \frac{V_j}{r_{ij}} \Theta_{ij} \hat{x}_{ij}}{1 + \sum_{j=1, j\neq i}^{N} \frac{V_j}{r_{ij}} \Theta_{ij}} \right| \tag{3.3}$$

Actually, an additional weighting factor $\frac{V_j}{r_{ij}}$, i.e. the volume of the neighbouring atom V_j divided by the distance r_{ij}, was introduced, since it was found to improve the correlation between the values of B_i and atomic solvation energies calculated by fdP (see next section). The value of B_i is normalised. It from 0 to 1 ($B_i = 0$ in case of a fully symmetric distribution and $B_i = 1$ in case of a fully asymmetric distribution). The additive constant of 1 in the denominator of Eq. 3.5 prevents the denominator from becoming zero for a completely isolated atom.

3.2.3 The degree of burial C

Once A and B are defined, it is possible to build a quadratic combination of them:

$$C_i = A_i + a_4 \cdot B_i + a_5 \cdot A_i \cdot B_i \tag{3.4}$$

where a_4 and a_5 are two parameters. This combination accounts for the volume, symmetry and a cross term and can be related to solvation energy by means of an

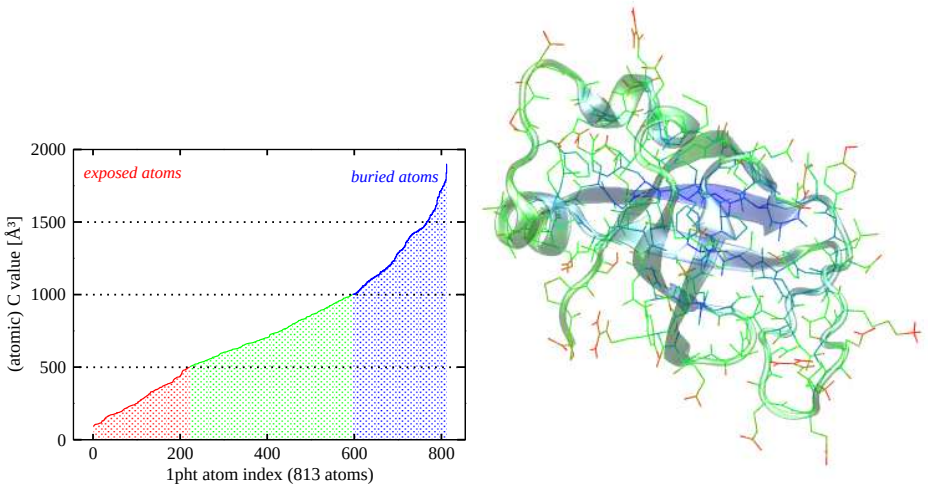


Figure 3.3: Protein 1pht atoms coloured according to their C value (red: $C = 0 \div 500$; green: $C = 500 \div 1000$; blue: $C = 1000 \div 1500$): the plot (left) shows the calculated C value for each atom (index in the x axis); the picture (left, stick and cartoon representations) shows that those atoms belonging to internal part of the protein are blue, while completely exposed atoms (in particular the ones belonging to sidechains extremities) are red. This is in agreement with the definition of C as a measure of the *degree of burial* of atoms.

empirical relationship between A, B and the atomic fdP. Several functional forms were tested, and the following quasi-sigmoidal function was chosen as it performed best:

$$\Delta G_i^{el,FACTS}(C_i) = a_0 + \frac{a_1}{1 + e^{a_2 \cdot (C_i + a3)}} = a_0 + \frac{a_1}{1 + e^{a_2 \cdot (a_3 + A_i + a_4 \cdot B_i + a_5 \cdot A_i \cdot B_i)}}. \qquad (3.5)$$

This choice will be justified by its good fit with fdP energies. $C_i = A_i + a_4 \cdot B_i + a_5 \cdot A_i \cdot B_i$ is then a measure of the solvent displacement around atom i or, conversely, a measure of the degree of burial of the atom within the internal dielectric, i.e. the molecule (the more an atom is buried, the greater is C_i). See Fig. 3.3 for details.

3.3 Use of A_i and B_i to derive $\Delta G_i^{solv,el}$

The previous section was devoted to an intuitive description of the link between the geometric meaning of A and B and the water encumbrance within the R^s sphere. Now we will move from intuition to physics or, more precisely, to statistics.

The most precise way to calculate the solvation properties of a continuum model is

to solve the aforementioned Poisson equation; in a GB model (like FACTS), in order to get the total solvation energy, it is necessary to calculate the solvation energy due to each atom, or atomic solvation energy ΔG_i^{solv}. Since A_i and B_i are somehow related to the atomic solvation features, it is interesting to plot the relationship between these atomic solvation energies and the related volume and symmetry measure for a certain molecule (see Fig. 3.5- 3.11).

Calculation of A, B parameters

The FACTS routine, by means of the keywords TPSR or TPSL, allows the user to print the couple (A, B) in the output file of CHARMM. Here follows an example:

```
FCTPRT> Atomic solvation energies:
        Atom number FACTS index Volume [A]   Symmetry [B] C            Solv. energy  Polar contr.
FCTSLF: 1           1           0.245752E+03 0.832140E+00 0.933070E+02 -0.639167E+02 -0.639167E+02
FCTSLF: 2           1           0.227268E+03 0.797356E+00 0.101048E+03 -0.626476E+02 -0.626476E+02
FCTSLF: 3           2           0.336225E+03 0.703686E+00 0.187106E+03 -0.412391E+02 -0.412391E+02
FCTSLF: 4           1           0.306768E+03 0.672609E+00 0.128267E+03 -0.583495E+02 -0.583495E+02
...     ...         ...         ...          ...          ...          ...           ...
```

where the `Atom number` column is related to the PDB atom numbering. The `FACTS index` field refers to one of the 7 atom types classified by the vdW/R^s criterion (for instance, FACTS index "1" corresponds to the atom type H*). The quantity `C` (and the remaining 2 fields) will be discussed below.

3.3.1 Calculation of the fdP atomic energies

The CHARMM command SOLV is invoked to compute the atomic solvation energies of each atom: the grid spacing in the finite-difference integration has been set to 0.2 Å; the distance between a protein atom and the edge has been set to 5 Å; the solvent dielectric constant (ϵ_s) has been set to 78.5, while the dielectric constant for the protein interior (ϵ_m) has been set to 2. Here follows the simple CHARMM input script designed to extract solvation energies from a trajectory snapshot.

```
! FDP ATOMIC SOLVATION ENERGY CALCULATIONS
```

```
SET EPSP   =   2                 ! DIELECTRIC CONSTANT FOR THE PROTEIN INTERIOR
SET EPSW   =   78.5              ! SOLVENT DIELECTRIC CONSTANT
SET CONC   =   0                 ! SALT CONCENTRATION
SET DCELC  =   0.2               ! THE GRID SPACING IN THE FINITE-DIFFERENCE
SET LEDGE  =   5                 ! DISTANCE BETWEEN A PROTEIN ATOM AND A GRID
SET OPTIONS =  WATR 1.4 REENTRANT ! USE OF THE MOLECULAR SURFACE
! GRID INFORMATION
COOR STAT
CALC XCEN = ( ?XMAX + ?XMIN ) / 2.0
CALC YCEN = ( ?YMAX + ?YMIN ) / 2.0
CALC ZCEN = ( ?ZMAX + ?ZMIN ) / 2.0
CALC NCLXC = INT ( ( @LEDGE * 4.0 + ?XMAX - ?XMIN ) / @DCELC )
CALC NCLYC = INT ( ( @LEDGE * 4.0 + ?YMAX - ?YMIN ) / @DCELC )
CALC NCLZC = INT ( ( @LEDGE * 4.0 + ?ZMAX - ?ZMIN ) / @DCELC )
SCALAR CHARGE SET 0.0 SELE ALL END     ! SWITCH OFF AL THE CHARGES
SCALAR CHARGE SET 1.0 SELE BYNU @N END ! SET THE CHARGE OF ATOM N TO 1
SCALAR WMAIN = RADIUS
SCALAR WMAIN SET 1.0 SELE TYPE H* END
PBEQ
SOLVE NCLX @NCLXC NCLY @NCLYC NCLZ @NCLZC DCEL @DCELC -
      EPSW @EPSP CONC @CONC INTBP @OPTIONS -
      XBCEN @XCEN  YBCEN @YCEN  ZBCEN @ZCEN
SET VACU ?ENPB             ! SOLVATION ENERGY OF ATOM @N IN VACUO
SOLVE NCLX @NCLXC NCLY @NCLYC NCLZ @NCLZC DCEL @DCELC -
      EPSW @EPSW CONC @CONC INTBP @OPTIONS -
      XBCEN @XCEN  YBCEN @YCEN  ZBCEN @ZCEN
SET WATER ?ENPB            ! SOLVATION ENERGY OF ATOM @N IN WATER
CALC SOLV = @WATER - @VACU ! ATOMIC SOLVATION FREE ENERGY OF ATOM @N
RESET
```

The software returns an approximation of the electrostatic atomic solvation free energy for each atom $\Delta G_{fdP,i}^{solv,el}$ (or just fdP from now on). The relationship between fdP and the geometric measures has to be setup empirically. In order to explore this empirical relation we should, ideally, compare volume, symmetry and atomic solvation energies of *all* the atoms in *all* the proteins, but this task is obviously not feasible. We must approximate the definition of such a (A, B, fdP) distribution with a finite testcase. It should be a large collection of molecular conformations that covers a wide range of "reachable" secondary structures, degree of burial and associated volume/symmetry pairs.

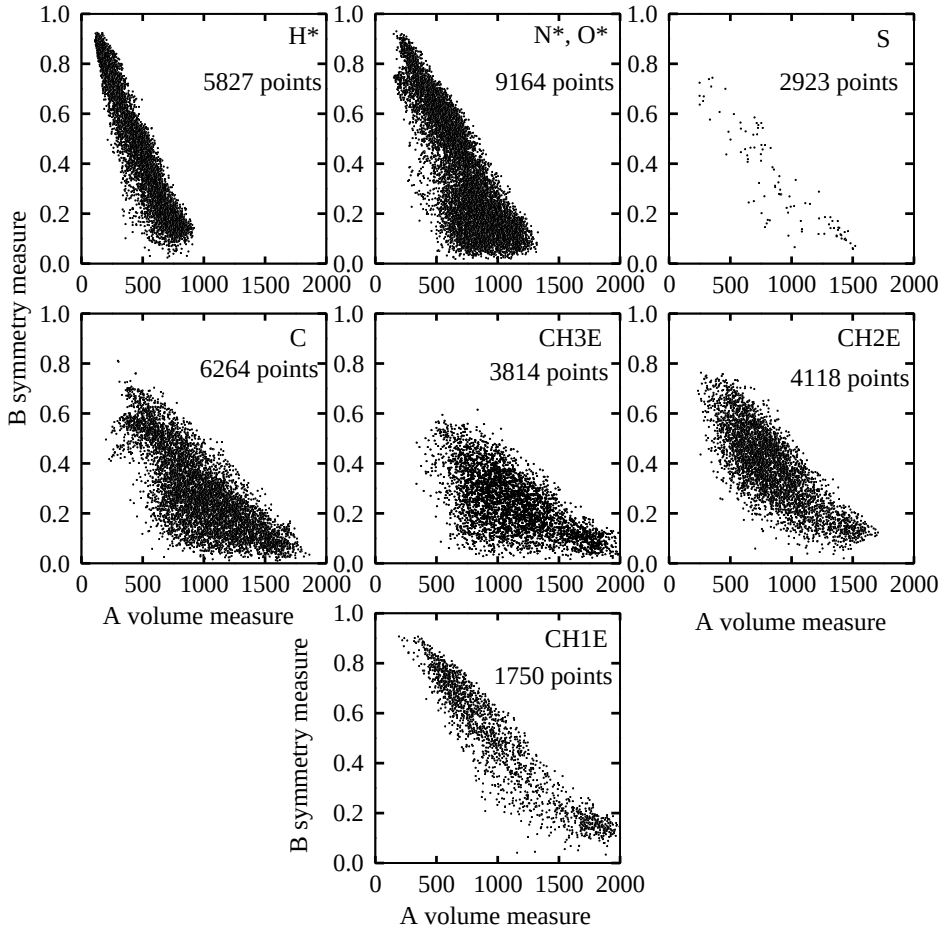


Figure 3.4: Distribution of A and B used in FACTS parametrisation (CHARMM param. 19, $\epsilon = 2$). The number of points is different according to the different amount of atom types in the proteins (indeed, solphure atoms are few sampled with respect to others). In general all the distributions are not uniformly sampled. Note the anticorrleation between the volume and symmetry measures. This is due to the fact that the more an atom is buried, the more the distribution of its surrounding atoms is likely to be uniform and, thus, symmetric. The more an atom is close to the surface, the more the distribution is likely to be asymmetric (only a part of the surrounding space around the selected atom is occupied by other atoms).

3.3.2 Fitting the (A,B, fdP) distributions

The original version of FACTS provide a testacase of 36 structures, each of them in 1-100 different configurations (including, beside the native state, a set of successive snapshots along thermal unfolding simulations, for a total of 1082 structures). The proteins involved in such a study were 1a2p, 1bpi, 1crn, 1dvd, 1f8a, 1hdn, 1inc, 1l2y, 1pgb, 1shg, 1ycq, 2a3d, 2ins, 3app, 5hvp, bet1, hlx1, prph, 1abz, 1cb3, 1cus, 1enh, 1fmk, 1hel, 1kvd, 1lz1, 1pht, 1ubq, 1ycr, 2ci2, 2ptl, 3pte, anki, gsgs, ins2, for a total of 5827 (A, B, fdP) points for atom type H*, 9164 for N* and O*, 2923 for S, 6224 for C, 1750 for CH1E, 4118 for CH2E and 3814 for CH3E. Fig. 3.4 summarises the results of this sampling.

Once the knowledge of A_i, B_i and fdP is at our disposal, it becomes possible to study the relationship between these quantities. This relationship could be recovered from this triplet (considering A and B as independent variables and fdP as dependent) via a *fitting procedure*. We have to formulate some hypothesis about the trend and than calculate the best fit via a parameter-optimisation procedure (*particle swarm* optimisation[4]. The fitting function chosen to represent the dependency between the geometric measures and the fdP atomic energies is the following:

$$f(A,B) = a_0 + \frac{a_1}{1 + e^{a_2 \cdot (a_3 + A + a_4 \cdot B + a_5 \cdot A \cdot B)}}, \tag{3.6}$$

which is very similar to a 2D-sigmoidal function. Fig. 3.5- 3.11 show the fit of such equation to the distribution related to the 7 FACTS indices. Once we have chosen a fitting function, we do not need the fdP calculations and can recover the solvation energy $\Delta G_{fit,i}^{solv,el}$ via this function. The atomic solvation energy calculation by means of the fitting functions (provided A and B for a given atom configuration) is a factor hundred faster than fdP calculations.

The whole method is effective because the time needed to calculate all the A_i and B_i is very short (to give an idea, the time needed to calculate via fdP – with an integration grid of 0.2 Å – all the $\Delta G_{fdP,i}^{solv,el}$ atoms of protein G is a matter of *hours*, while the time needed to perform A, B and $\Delta G_{fit,i}^{solv,el}$ calculation is a matter of fractions of *seconds*).

[4]See Eberhart and Hu, *Human tremor analysis using particle swarm optimisation*, Proceedings of Congress on Evolutionary Computation, 1927-1930 (1999) and the following websites: http://www.particleswarm.info/ and http://www.swarmintelligence.org

The drawback lies in the **accuracy** by which the fitted values are calculated. The analysis of the percentage error distribution between the fit and the dataset involved in the fitting process (see next chapter) shows that the biggest source of errors are the inner atoms, meaning that even if the average error is around 10-20% between fitted and fdP data, the more an atom is buried, the more FACTS fails in deriving its solvation energy. Buried atoms have a small ΔG^{solv}, which means that the absolute value of the error should be small. But they are also *numerous*, as one can easily argue from Fig. 3.3. This, and other problems related to the error analysis of the FACTS parametrisation, will be discussed in details in chapter 5.

3.3.3 Introduction to Manuscript 1

The aforementioned hypothesis of a systematic error in the FACTS electrostatics parametrisation on one hand and the need of a more complex determination of the nonpolar contribution (as we seen in chapter 2) on the other hand come from a **systematic study** of FACTS behaviour with proteins and peptides. The detailed analyses of such a study are presented in the following Manuscript 1.

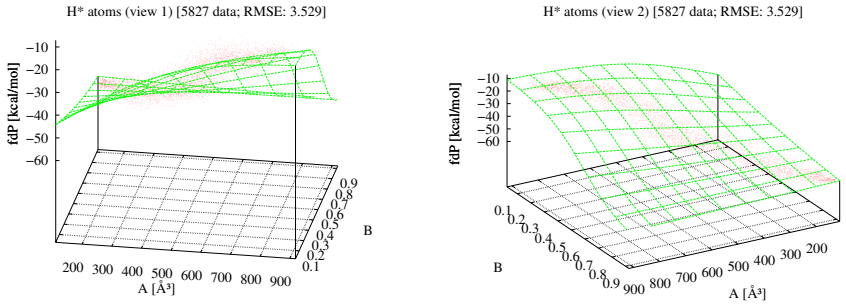


Figure 3.5: Testcase fitting (5827 data points, RMSE=3.529 [kcal/mol]) for H* atoms (same plot, two views).

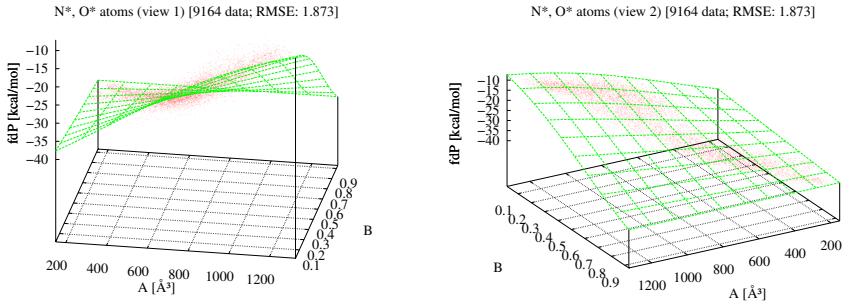


Figure 3.6: Testcase fitting (9164 data points, RMSE=1.873 [kcal/mol]) for N*, O* atoms (same plot, two views).

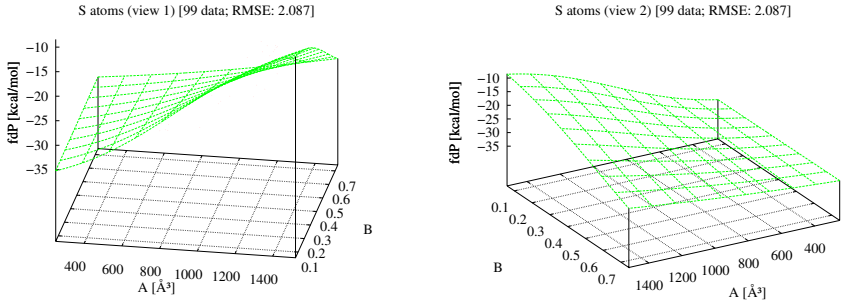


Figure 3.7: Testcase fitting (99 data points, RMSE=2.087 [kcal/mol]) for S atoms (same plot, two views).

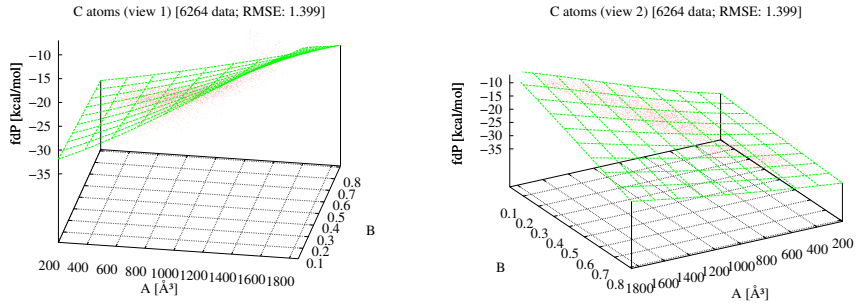


Figure 3.8: Testcase fitting (6264 data points, RMSE=1.399 [kcal/mol]) for C atoms (same plot, two views).

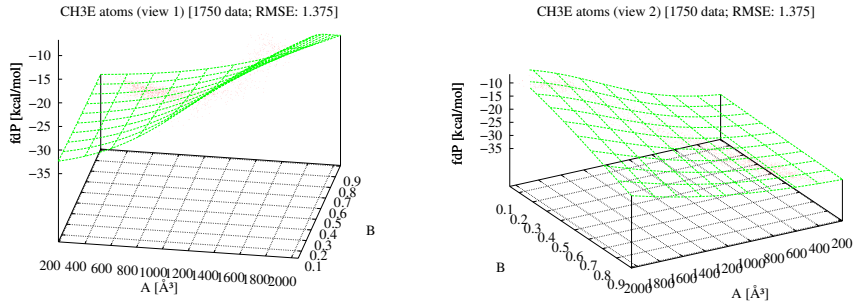


Figure 3.9: Testcase fitting (1750 data points, RMSE=1.375 [kcal/mol]) for CH3E atoms (same plot, two views).

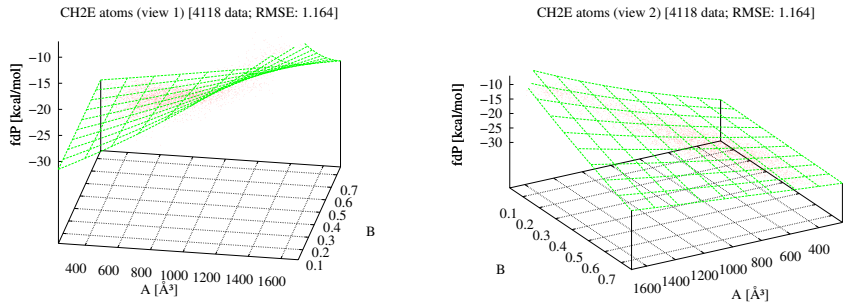


Figure 3.10: Testcase fitting (4118 data points, RMSE=1.164 [kcal/mol]) for CH2E atoms (same plot, two views).

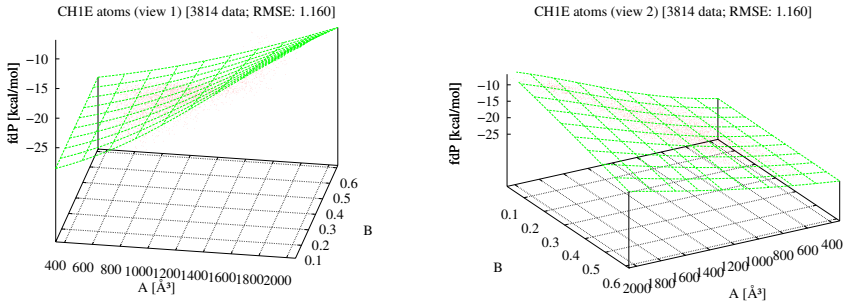


Figure 3.11: Testcase fitting (3814 data points, RMSE=1.160 [kcal/mol]) for CH1E atoms (same plot, two views).

Chapter 4

FACTS: A Systematic Study

This chapter contains the first part of my original contribution to the development of FACTS. Hundreds of MD simulations of different macromolecules were run in order to systematically test the FACTS features. This chapter is actually the outline of a paper which is going to be submitted (Supplementary Material and Figures can be found in Appendix 1).

ABSTRACT: The FACTS implicit solvent model is tested by using unstructured and structured peptides, as well as small globular proteins. Experimental data are used to systematically select the best FACTS parameters. Results indicate FACTS as a powerful tool to investigate the biological molecules by molecular dynamics within their aqueous environment, provided the assessment of a transferable set of parameters. Evidence that such a parametrisation is highly sensitive to the degree of structure of the biomolecule suggests that a more sophisticated treatment of nonpolar contribution of FACTS is crucial in order to implicitly reproduce the effect of water on biological molecules.

ABBREVIATIONS: MD: molecular dynamics; CS: chemical shifts.

4.1 Introduction

It is a challenge to accurately and efficiently approximate aqueous solvent effects on biological macromolecules molecular dynamics (MD) simulations. Many implicit solvent models have recently been developed [1], especially on the basis of the generalised Born approximation (GB) – for the treatment of electrostatics – and the solvent accessible surface area approach (SASA) – for the treatment of nonpolar interactions [2, 3, 4, 5].

These developments often involve a more refined nonpolar solvation energy treatment [6] which is a weakness of implicit models [7, 8, 9].

FACTS is a GB/SASA model [10]. It is based on the fully analytical calculation of the volume and spatial symmetry of the solvent displacement around each atom by their neighbours. This geometric information are used to approximate the self electrostatic solvation energy and the SASA. The former yields the effective Born radius which is needed to recover the solvent-screened electrostatic interaction energy, while the latter is used to determine the nonpolar solvation energy by the common (linear) SASA model for nonpolar interactions.

Here, a systematic study of FACTS behaviour with unstructured and reversible folding peptides is presented. Experimental data is used to assess the best parametrisation either for electrostatic and nonpolar interactions. Studies of peptide-like structures at equilibrium with FACTS show that electrostatic and nonpolar interactions should be adjusted according to the structure of the biomolecule under consideration.

The precise determination of Born radii is actually of extreme relevance to ensure an accountable assessment of the electrostatic interactions [11]. The solvation free energy in FACTS is written as the sum of a polar and a nonpolar term $\Delta G^{FACTS} = \Delta G^{el,FACTS} + \Delta G^{np,FACTS}$. The electrostatic solvation free energy of atom i, ΔG_i^{el}, is calculated by taking into account the volume A_i and the symmetry B_i of the distribution of neighbouring atoms within a sphere of radius R_i^{sphere} around the atom i itself. The relationship between atomic solvation energies $\Delta G_i^{el,FACTS}$ (for a unit charge) and the quantities A_i and B_i is then recovered by fitting a sigmoidal function whose parameters have been optimised to accurately reproduce the finite difference solutions of the Poisson equation taken as reference values:

$$\Delta G_i^{el,FACTS} = a_0 + \frac{a_1}{1 + e^{-a_2(A_i + b_1 B_i + b_2 A_i B_i - a_3)}},$$

where the quantity $C_i = A_i + b_1 B_i + b_2 A_i B_i$ is a measure of solvent displacement or, equivalently, of the degree of burial of the atom i inside the macromolecule (See Fig. 9.82). Now, by definition (and being q_i the charge of atom i and $\tau = 1/\varepsilon_{solute} - 1/\varepsilon_{solvent}$), we can write the FACTS effective Born radii $R_i^{FACTS} = -\tau q_i^2 / 2\Delta G_i^{el,FACTS}$. These radii are then used in the classic GB formula [12] to compute the *total* electro-

static solvation energy:

$$\Delta G^{el,FACTS} = -\frac{\tau}{2}\sum_{i,j=1}^{N}\frac{q_i q_j}{\sqrt{r_{ij}^2 + R_i^{FACTS}R_j^{FACTS}\exp(-r_{ij}^2/\kappa R_i^{FACTS}R_j^{FACTS})}} \quad (4.1)$$

where r_{ij} is the distance between charges q_i and q_j, $r_{ii} = 0$. The Still's constant κ is usually set to 4 or 8, and N is the number of atoms in the solute. The nonpolar solvation free energy of atom i, $\Delta G_i^{np,FACTS}$ is related to SASA as follows:

$$\Delta G_i^{np,FACTS} = \gamma \cdot S_i, \quad (4.2)$$

where S_i is the contribution of atom i to the total SASA and γ the coefficient (often called "surface tension") between the former and the nonpolar solvation energy, according to the SASA approximation (the *total* solvation energy $\Delta G^{np,FACTS}$ is simply the sum over all the $\Delta G_i^{np,FACTS}$). In order to evaluate S_i, using the FACTS geometric measures described above, and being $D_i = A_i + d_1 B_i + d_2 A_i B_i$ the combination of volume and symmetry to be parametrised over experimental SASA, one can define:

$$S_i^{FACTS} = c_0 + \frac{c_1}{1 + e^{-c_2(D_i - c_3)}},$$

The parameters d_1, d_2, c_0, c_1 c_2 and c_3 are optimised by fitting them to the atomic surface areas obtained by an analytic method. In a previous work, MD simulations results showed that the native state of structured peptides and proteins were stable using FACTS, while marginally stable peptides and unstructured loops in proteins were flexible [10].

4.2 Methods

4.2.1 FACTS parameters

The FACTS parameters which can be changed by the user (free parameters) are ϵ_{solute}, which will be simply referred to as ϵ from now on (while $\epsilon_{solvent} = 78.5$ is constant), and the surface tension γ (units: cal·mol^{-}1·Å^2) will be dropped. The more ϵ is close to unity, the more the internal and external dielectric constants are different: increasing ϵ, thus, means to weaken the electrostatic interactions within the macromolecules with respect to the interactions between these atoms and the solvent dielectrics. On the other hand, increasing γ will result in a stronger nonpolar interaction, and thus a more

compact molecule. Here, four combination of internal dielectric and surface tension were tested in order to mix up these different tendencies of the free parameters to stabilise/destabilise the polypeptide structures:

($\epsilon = 1$, $\gamma = 7.5$) = **I** : strong (internal) electrostatic interactions, weak nonpolar interactions; expected to reduce the stability of globular structures, to keep unstructured peptides elongated and to slow down kinetics.

($\epsilon = 1$, $\gamma = 15$) = **II** : strong electrostatics, strong nonpolar interactions.

($\epsilon = 2$, $\gamma = 7.5$) = **III** : weak electrostatics, weak nonpolar interactions; setting $\epsilon = 2$ means to halve (in module) the electrostatics contribution to the solvation energy, since $\tau = \frac{78.5-2}{78.5 \cdot 2} \simeq 0.487$ while $\tau = \frac{78.5-1}{78.5 \cdot 1} \simeq 0.988$ in eq. 4.1.

($\epsilon = 2$, $\gamma = 15$) = **IV** : weak electrostatics, strong nonpolar interactions.

Since the purpose of a solvation model is to provide a *unique* parametrisation, the first target of this work is to find the *best* among these I, II, II or IV FACTS parameter sets (internal dielectric and surface tension).

MD simulation shows that the model acts ambiguously toward the solvation properties of unstructured and structured peptides, meaning that best results are attained by *different* parameter sets. This suggests the need for a correction, which takes into account the different organisation of the polypeptide structures.

4.2.2 MD simulations with CHARMM

The MD simulations were performed by the CHARMM program with the polar hydrogen parameter set PARAM19 [13] and using leap-frog integration with a Berendsen thermostat (coupling constant: 5 ps; velocities were reassigned every 20-40 ns), always with an integration time step of 2 fs. The CHARMM default truncation scheme of long-range electrostatics and van der Waals energy was used (SHIFT to 0 energy at 7.5 Å). The non-bonding interactions were updated heuristically. The SHAKE algorithm was used. Coordinate frames were saved every 20 ps.

4.2.3 Testing convergence of peptides simulations

Before doing any kind of analysis on the MD simulations, preliminary convergence tests were performed to assess whether each peptide's conformational space was properly

sampled. Convergence of computer simulations is a critical subject, due to theoretical [14] and computational issues [15, 16, 17]. Actually, a definitive solution to this problem is not currently available. For an intriguing glance on the complexity of this problem, see the discussion on *Angewandte Chemie* between van Gunsteren and coworkers and Karplus and coworkers [18, 19, 20]. Here we adopted two different approaches to investigate sampling features. The first method consists in halving the trajectories and calculating the root mean square deviation with respect to Cα atoms (RMSD) of a reference structure and the radius of gyration (RGYR) distributions for each part. The percent deviation between these distributions is then referred to as a first convergence measure (the MD simulations in this paper give rise to deviations which do not exceed 5%).

The second method consists in recovering the time series of the number of *significantly populated* conformation clusters. The trajectories have been divided in 100 sections. The number of clusters of the first section has been calculated (by means of the Leader algorithm implemented in Wordom [21] with a cutoff of 2.5 Å). The threshold needed to consider a cluster sufficiently populated has been set to 0.5%. An analogous clustering has been performed for the first section together with the second, giving a second measure of significantly populated clusters (which is usually greater than the previous one). Since the studied polypeptide structures range between 12 and 36 residues, 2 μs are usually sufficient for reaching a plateau in these time series (convergence tests related to each MD study and involving either RMSD and RGYR deviations and clusterings are systematically reported in the Supplementary Material). Reaching such a plateau is taken as the second indication that the MD simulation had sufficiently explored its conformational space [22, 23].

4.2.4 Use of chemical shifts to assess FACTS features with structured peptides

The SHIFTX program [24] has been used to recover δHCα, δHCβ, δHN and δCα NMR chemical shifts (CS) starting from the CHARMM trajectories with FACTS related to structured peptides. The software returns a set of CS starting from a selected pdb file, corresponding to a selected frame from trajectories. The correlation between experi-

mental and calculated shifts is of about 80-90% either for ^{1}H ^{1}3C and ^{1}5N prediction, whereas the error in their estimation is lower for δHCα and δHCβ (5-10%) and higher for δHN and δCα (20-30%).

In order to evaluate a CS it is required to average all the CS obtained for each conformation explored by the MD simulation [25, 26, 27, 28]. These average values can be compared with experimental peaks provided that MD simulations have reached convergence. Although SHIFTX has been parametrised using folded structures, the (averaged) CS related to the MD simulations (which actually are in equilibrium between folded and unfolded state) are still valid, since it has been shown that the non-structured fragments of the proteins used to parametrise SHIFTX are a good model for the prediction of random-coil CS [29]. Nevertheless, in this work SHIFTX has never been used for CS prediction of unstructured peptides. See Supplementary Material for more details about the statistical analysis of the CS evaluation.

4.3 Results and Discussion

In this section we present the FACTS model using unstructured and structured peptides (a fragment of tyrosine hydroxylase, the monomeric melittin, a β-hairpin, an α-helix and a three stranded, antiparallel β-sheet). Several quantities are measured and compared to experimental data. An asterix (*) will score the FACTS parametrisation(s) among I, II, III and IV indicating the best result in each test.

4.3.1 Overview of FACTS electrostatics setup

Before conducting extensive MD simulations, we explored the electrostatics features of FACTS. Lazaridis and coworkers [30], showed that an useful tool to assess the electrostatics parameters for a solvation model is the evaluation of the potentials of mean force (PMF) of ionisable sidechains by comparing it with the ones for explicit water (EW). In Supplementary Material the results with FACTS are reported for the different parametrisations. Here, as a relevant example, the results related to the arginine-glutamic acid sidechain analogues in collinear approach are presented in Fig. 9.83 and summarised in Tab 4.1. The PMFs show that FACTS III and IV,

source	$\lvert\Delta^{EW}\rvert$ [kcal/mol]
GBMV	0.60
GBSW	1.96
EEF1	4.21
FACTS I	7.56
FACTS II	7.44
FACTS III*	2.96
FACTS IV	3.82

Table 4.1: Difference between the minimum of the arginine-glutamic acid PMF (collinear approach) obtained by Lazaridis and coworkers EW simulations ($\simeq -4.5$ kcal/mol) and the ones obtained with implicit solvent models (see Fig. 9.83). FACTS III is the parameter set which better approximates the EW profile.

according to the difference between the PMF minimum obtained by FACTS and the one calculated via EW, behave as EEF1 while FACTS I and II give more prominent minima, in agreement with the higher electrostatic strength related to the $\epsilon = 1$ set up.

4.3.2 Unstructured peptides: Tyrosine hydroxylase (22-34)

Tyrosine hydroxylase is the enzyme responsible for catalyzing the conversion of the amino acid L-tyrosine to dihydroxy phenylalanine. Fluorescent-resonance-energy-transfer (FRET) and MD studies related to the end-to-end distance of the fragment 22-34 of this enzyme were performed by Stultz and coworkers [31, 32]. FACTS attained very interesting results with the PMF of the fragment W-KQAEAVTSPR-W (two additional tryptophan residue were located at the peptide terminus to mimic the acceptor and donor groups used in FRET experiments). A single simulation of 4 μs for each FACTS setup was performed (convergence tests reported in Supplementary Material).

FRET measurements: The PMF for folding the 12 residue peptide from an extended to a compact state (meaning that peptide extremities are in close proximity) was calculated by means of the relative frequency $f(r)$ of the end-to-end distance r along the trajectory. In particular $PMF(r) = -k_B \cdot T \cdot \ln\ f(r)$, where r is the distance

between the Cα atoms of each tryptophan. The FRET efficiency was computed from PMF and compared to the experimental result [31]. With a Förster critical distance R_0 equal to 23.6 Å and being $E(r) = R_0^6/(R_0^6 + r^6)$ the statistical mechanical expression for the FRET efficiency for peptides E [33], it can be shown [34] that the latter is related to PMF as follows:

$$E \simeq \frac{\int_m^M E(r) e^{-PMF(r)/kT} dr}{\int_m^M e^{-PMF(r)/kT} dr}, \tag{4.3}$$

where m and M are the minimum and maximum value of r. Simulations performed with $\epsilon = 2$ are consistent with other implicit solvent models. The ones related to $\epsilon = 1$ are in close agreement with experimental data, namely for FACTS I. Four additional simulations with FACTS III were performed in order to estimate the uncertainty on E, which results of about 0.02 units.

source	FRET	⟨W-W dist.⟩ [Å]
exp.	**0.46**	-
EW	0.50	*22.9*
ACE	0.99	-
EEF1	1.00	-
GBMV	0.97	-
SASA	1.00	-
FACTS I	0.31(2)	*27.7*
FACTS II*	0.44(2)	*24.7*
FACTS III	0.83(2)	*15.0*
FACTS IV	0.97(2)	*10.3*

Table 4.2: (Central column) FRET calculations related to different solvation models and FACTS, obtained using PMFs (reported in Supplementary Material) in comparison with experimental data (bold font). Rightmost column shows EW data (italic) related to the fragment end-to-end distance. FACTS II (related to $\epsilon = 1$) attains a remarkable match either with explicit and experimental data.

Comparison with EW: We compared the energy landscape and the average distance between the tryptophans' Cα atoms obtained with FACTS and the ones obtained by Stultz via EW calculations. Results are summarised in Tab. 4.2 (see Supplementary Material for PMF plots). FACTS I and II reveal interesting matches with EW data.

4.3.3 Unstructured peptides: Melittin

High resolution ^{1}H-NMR studies of monomeric melittin (GIGAVLKVLTTGLPALISWIKRKRQQ) in aqueous solution (at 360 MHz, pH 3.0 and 30°C) showed that such a polypeptide is predominantly in an unstructured and flexible form [35, 36], mostly unstructured [37] and with a low helix content [38]. Four simulations of 6 μs each were performed with FACTS starting from an extended and equilibrated structure with different internal dielectrics/surface tension parameters.

Secondary structure. Secondary structure analysis of the conformations is presented in Tab. 6.2. FACTS II favours the β-strand formation between residues 1-10 and 11-24 up to 53% of the entire trajectory. FACTS IV favours the α-helix formation between residues 13-26 up to 22%. This is in contrast with the hypothesis of an unstructured peptide. Nevertheless, both the simulations related to FACTS I and FACTS III are closer to the random-coil hypothesis [37].

sec. str./FACTS	I*	II	III	IV	EW	ORD
β strand	0.02	0.53	0.08	0.11	-	**0.00**
3-10 helix	0.06	0.10	0.10	0.11	*0.02*	-
bend	0.00	0.01	0.01	0.02	*0.15*	-
turn	0.03	0.04	0.12	0.10	*0.09*	-
random-coil	0.88	0.32	0.49	0.39	*0.29*	**0.88**
α helix	0.01	0.00	0.16	0.22	*0.42*	**0.12**
π helix	0.00	0.00	0.03	0.05	*0.03*	-

Table 4.3: Secondary structure analysis of melittin with FACTS, performed with DSSP [39] program. Comparison with EW simulations [40] (italic) and optical rotary dispersion (ORD) [41]. Notice the decreasing random-coil percentage as a consequence of internal dielectric and/or surface tension increasing. π-helix, short β-bridges, turns, bends and 3-10 helix, are highly ephemeral along the trajectories and not specific to any residue. FACTS I resulted as the best parametrisation choice with respect to the ability of the model to reproduce melittin random-coil percentage in water.

Fluorescent energy transfer measurements: The distribution of distance between Cγ1 of Val 5 and Cγ of Trp 19 was compared with the one evaluated by fluorescent energy

transfer measurements on melittin mutant V5Y in sodium phosphate buffer, pH 7.4 [38]. Results are reported in Fig. 9.84. FACTS I is more accurate, since it shows a unimodal distribution (in agreement with random-coil hypothesis, as quoted in [38]) and a maximum which is closer to experimental data, compared to the other parametrisations. In particular FACTS I* shows a maximum at 36.95 Å, closer to the experimental value of $\simeq 30 - 32$ Å, while FACTS II shows a maximum in the distribution (not unimodal) at 10.02 Å, FACTS III at 11.87 (unimodal) and FACTS IV at 10.17 (unimodal).

4.3.4 Structured peptides: β-hairpin of protein G

The FACTS behaviour with a simple β-hairpin was investigated. The fragment is derived from the B1 domain of the streptococcal energy protein (1igd), and it contains the only natural sequence which folds in a native-like β-hairpin structure in water [42, 43].

Comparison with CS at 278 K. Blanco and coworkers [42] studied ^{1}H CS of the 41-56 fragment of B1 domain of protein G at 278 K, 5 nM sodium phosphate and pH = 6.3, $H_2O/^2H_2O$ (9:1 by vol.). The complete list of calculated CS (δHN, δHCα and δHCβ for each parameter set at 280 K) as well as the analysis of the best FACTS setup, is reported in Tab. 9.4, 9.6 and 9.8. The comparison with Blanco's experimental CS gives ambiguous results, likely because of the difficulty of reaching convergence at this temperature (see Supplementary Material).

par.	DF	χ^2	p
FACTS I	16	16.066	0.250<p<0.500
FACTS II	16	10.2157	0.750<p<0.900
FACTS III*	16	7.58954	0.950<p<0.975
FACTS IV*	16	7.35341	0.950<p<0.975

Table 4.4: Statistical analysis of pgbh HCα shifts (at 280 K) coming from the comparison of 16 experimental/calculated values. The stars indicate the best fit (see Supp. Mat. for details).

Comparison with δHCα trends from 278 to 338 K: In order to test the FACTS be-

par.	DF	χ^2	p
FACTS I	17	29.8503	0.025<p<0.050
FACTS II*	17	25.3562	0.050<p<0.100
FACTS III	17	39.2228	p<0.005
FACTS IV	17	31.5028	0.010<p<0.025

Table 4.5: Statistical analysis of pgbh HCβ shifts (at 280 K): coming from the comparison of 17 experimental/calculated values. The star indicates the best fit (see Supp. Mat. for details).

par.	DF	χ^2	p
FACTS I	15	25.6522	0.025<p<0.050
FACTS II	15	42.4593	p<0.005
FACTS III*	15	4.35024	0.995<p<0.999
FACTS IV	15	7.60429	0.900<p<0.950

Table 4.6: Statistical analysis of pgbh HN shifts (at 280 K): coming from the comparison of 15 experimental/calculated values. The star indicates the best fit (see Supp. Mat. for details).

haviour with the simple hairpin, a higher temperature experiment is useful. Honda and coworkers [43] studied the thermodynamic properties of the pgbh β-hairpin formation via NMR melting measurements with high accuracy. In particular, the temperature dependence of δHCα in 99.996 % ^2H_2O with 5 nM sodium phosphate buffer (p^{2}H 7.0) between 278 and 338 K has been investigated. By the accurate determination of the inflection point of such a dependence it is possible to recover the molar fraction of the unfolded molecule. Indeed, with the assumption of a constant heat capacity (which is the case of pgbh, since its SASA is very small, see for instance [44]), and a two state transition, the fraction of unfolded structures f is simply related to the equilibrium constant k by means of $f = k/(1+k)$, and $k = e^{-\frac{\Delta H_m}{R}\left(\frac{1}{T}-\frac{1}{T_m}\right)}$, where T_m is the inflection point temperature of the melting curve and ΔH_m is the change in enthalpy upon unfolding at the transition temperature T_m. We performed 4 μs long simulations of pgbh with FACTS I, II, III, and IV at 270, 280, 290...350 K and for most of the temperatures convergence is likely to be satisfactory (see Supplementary Material). The Cα-proton CS for each residue has been extracted. Then, their relations with the temperature have been fitted to a sigmoidal function in the form $\Gamma = \Gamma_F + (\Gamma_U - \Gamma_F)\cdot f$

(see Supplementary Material for details about this study). Parameters were averaged, leading to the melting curve in Fig. 9.85. The analysis is supported by the study of the percentage of unfolded conformation along the MD simulation on the basis of the RMSD. Here the peptide is considered unfolded when the RMSD of a conformation with respect to 1pga (41-56) exceed 2.5 Å.

Both of the analyses, as far as wkqa and melittin are concerned, reveal FACTS I and II to be closer to experimental data than FACTS III and IV. In particular, the latter show a weaker response to the temperature variation. But this effect is due to the lower magnitude of the electrostatic contribution. However, the nonpolar contribution does not affect too much the dynamics (in Fig. 9.85, FACTS I-II and FACTS III-IV either according to the CS and the RMSD analysis behave quite similarly). Thus, despite a suspicion about convergence at low temperature, pgbh features are not well reproduced with the selected parametrisations – as in the case of wkqa and melittin. It is important to point out that the difference between the former and the latter is mainly in the intrinsic tendency of pgbh to form secondary structure, suggesting the hypothesis of a relationship between this propensity and the choice of the right FACTS parametrisation.

4.3.5 Structured peptides: Ac-$(AAQAA)_3$-NH_2 helical peptide

In their work, Stellwagen and coworkers measured the carbonyl-carbon CS thermal dependence of the peptide Ac-$(AAQAA)_3$-NH_2 at various temperature, deriving the residue distribution of helical content at 0°C, pH 7 [45]. In order to investigate FACTS characteristics with this model of helical peptide, four simulations (with FACTS I, II, II, IV) of 4 μs each with at 274 K with FACTS (starting from an extended act2 structure) were performed (convergence tests are reported in Supplementary Material).

Comparison with carbonyl-carbon chemical shifts: An accurate analysis of the CS obtained by averaging the carbonyl-carbon shifts from the MD simulations allows to clearly identify the best parameter set, at least for the value of the internal dielectrics. See Tab. 4.7 and Tab. 4.8. FACTS II and IV perform clearly better in this case than FACTS I and II, related to $\epsilon = 1$ (see also Fig. 9.87). This is exactly the opposite case

of unstructured peptides wkqa and melittin.

par.	DF	χ^2	p
FACTS I	10	127.36	$p < 0.005$
FACTS II	10	113.2	$p < 0.005$
FACTS III*	10	17.2028	$0.050 < p < 0.100$
FACTS IV*	10	17.3573	$0.050 < p < 0.100$

Table 4.7: Statistical analysis of act2 δC-*helix* shifts extracted from FACTS trajectories in comparison with 10 experimental resonances. The experimental shifts are related to helical conformation (274 K): FACTS III and IV allow the peptide to visit α-helical conformations, in contrast with FACTS I and II.

par.	DF	χ^2	p
FACTS I	14	26.4736	$0.010 < p < 0.025$
FACTS II	14	25.4593	$0.025 < p < 0.050$
FACTS III*	14	340.438	$p < 0.005$
FACTS IV*	14	220.806	$p < 0.005$

Table 4.8: Statistical analysis of act2 δC-*coil* shifts. The analysis show the peptide being most of the time in a random coil conformation with FACTS I and II. The chemical shifts used for this counterproof are related to $T = 90$°C as in Ref. [45], in which the peptide is mostly unstructured as revealed by circular dichroic measurements.

Distribution of helicity: Experimental value of the helicity fraction h along the peptide chain provided by Stellwagen and coworkers are recovered via CS measurements by $h = \frac{\delta-\delta_c}{\delta_\alpha-\delta_c}$ [46], being δ, δ_α and δ_c the observed (carbonyl-carbon) shift, the chemical shift of the helix conformation and the chemical shift of the coil conformation, respectively. Errors for the experimental helicity are recovered via gaussian propagation on the CS standard deviation (which is $\simeq 0.07$ ppm) using the previous expression for helicity and evaluated around 0.1 units.

However, a single residue is normally considered helical if it belongs to a segment of at least 3 residues whose dihedral angles differ less than 30°C from the nominal values $\phi = -57$°C and $\psi = -47$°C [47]. In Fig. 9.88 we plot the helicity calculated

with both methods (CS based and three-segment based) together with a comparison with other implicit models. Note the helicity per residue calculated by means of CS recovered with SHIFTX program is in agreement with the one calculated via dihedral angles. Moreover, as we can expect from CS analysis, FACTS I and II fail to reproduce the experimental profile, while FACTS III* and IV* are as close experimental as SCP model [48] and more realistic than SASA [49].

Interestingly, Stellenwagen and coworkers showed that lowering the pH from 7 to 2 in absence of NaCl results in decreasing the shifts, suggesting that "the stability of the peptide is not due to electrostatics (experimental) interactions". Therefore the big difference observed between FACTS I-II and FACTS III-IV should rather de due to the *assessment of the nonpolar contribution* than due to parametrisation (different values of surface tension, indeed, lead to similar results).

4.3.6 Structured peptides: Three-stranded β sheets

A three-stranded antiparallel β-sheet peptide (gsgs) made up by 20 amino acids TWI-QNGSTKWYQHGSTKIYT with experimental folding rate of μs, was used to stress FACTS towards reversible folding [50]. Simulations of 7 μs for each FACTS setup were performed at 300 K. Fig. 9.89 shows the time series of RMSD and native contacts with respect to Cα atoms (convergence tests are reported in the Supplementary Material).

Native state. The gsgs peptide was experimentally seen to contain predominately β-sheets. As a preliminary, qualitatively study of the FACTS "native" structure under the four different FACTS parametrisations, a cluster analysis of these simulations has been performed with the Leader algorithm by Wordom in order to recover the most populated conformations for each parameter set of FACTS I, II, III, IV. Cluster cutoff is 2.5 Å. The four different parametrisations of FACTS lead to the structures shown in Fig. 9.90.

CS analysis: Tab. 9.15, 9.17 and 9.19 show the comparison between experimental and calculated HCα, HCβ and HN shifts. FACTS behaves better with the III and IV parameter sets although also FACTS III gives reasonable results. Either on the basis

of cluster analysis (which show the peptide being mostly unstructured) and the CS analysis, the parameter set FACTS I have to, conversely, be rejected.

par.	DF	χ^2	p
FACTS I	19	15.6817	0.500<p<0.750
FACTS II	19	7.89895	0.975<p<0.990
FACTS III*	19	6.06333	0.995<p<0.999
FACTS IV*	19	5.57689	0.995<p<0.999

Table 4.9: Statistical analysis of gsgs HCα shifts (at 300 K) on the basis of a comparison between 19 experimental and calculated values. The stars indicate the best fit (see Supp. Mat. for details).

par.	DF	χ^2	p
FACTS I	28	182.434	p<0.005
FACTS II	28	27.6118	0.250<p<0.500
FACTS III	28	27.6917	0.250<p<0.500
FACTS IV*	28	22.137	0.750<p<0.900

Table 4.10: Statistical analysis of gsgs HCβ shifts (at 300 K) on the basis of a comparison between 28 experimental/calculated values. The star indicates the best fit (see Supp. Mat. for details).

par.	DF	χ^2	p
FACTS I	19	37.8765	0.005<p<0.010
FACTS II	19	13.765	0.750<p<0.900
FACTS III*	19	6.80317	0.995<p<0.999
FACTS* IV	19	7.21109	0.995<p<0.999

Table 4.11: Statistical analysis of gsgs HN shifts (at 300 K) on the basis of a comparison between 19 experimental/calculated values. The stars indicate the best fit (see Supp. Mat. for details).

NOEs violations analysis Tab. 6.4 sums up the HCα-HCα NOEs violations. The analysis of these quantities confirms the results coming from the CS analysis (FACTS II, III and IV give similar result and I has to be rejected). Interestingly, FACTS II and IV (which differ for electrostatics setup and have identical value of surface tension) behave quite similarly.

exp. NOEs/FACTS	I	II*	III	IV*
very weak	0	1	0	0
weak	4	3	3	3
medium	5	3	4	3
medium-strong	2	0	1	0
strong	0	0	0	0

Table 4.12: Violations of the medium- and long-range NOE connectivities of the gsgs peptide, related to FACTS with different parametrisations. Experimental data are related to 1 mM of gsgs peptide in aqueous solution, pH 3.4, at 10°C. Simulations are 6 μs long. See Table 1 of [50] and Supplementary Material for details. Notice that FACTS II and FACTS IV (different electrostatics set up) give (almost) the same number of violations.

Secondary structure: De Alba and coworkers, derived the population of β-sheets using the intensity of HCα-HCα NOEs [50, 51] between relevant residues. In particular they selected W2-H12 and Q4-K9 resonances to estimate the β-structure between the first and the central sheets while Q12-K17 and W10-Y19 resonances were used to estimate it for the central and the third sheets. In order to compare this estimation with MD results, we calculated the ratio between the average inter-β-sheet distances occurring between the selected HCα and the nominal proton-proton distance in a regular, non twisted, antiparallel β-structure [52], namely HCα-HCα $= 4.3 \pm 1.3$ Å(see Fig. 4.13).

	exp./FACTS	I	II	III*	IV*
W2-Y11	**0.24**	0.06	0.24	0.16	0.20
Q4-K9	**0.31**	0.04	0.05	0.16	0.16
Q12-K17	**0.19**	0.02	0.04	0.12	0.13
W10-Y19	**0.13**	0.05	0.22	0.23	0.30

Table 4.13: Estimation of the percent population of β-structures formed by gsgs between W2-Y11, Q4-K9, Q12-K17 and W10-Y19 with FACTS at 300 K, according to the different parameter sets. Comparison with experimental values calculated from the HCα-HCα NOE intensity at pH 3.25. FACTS III and IV give globally the more reliable results (while the overall discrepancy from De Alba estimation of simulations with FACTS I and II is almost double than the one coming from simulations with $\epsilon = 2$) even if they favour the proton-proton distance related to β-structure more between the central and third sheets instead of the one between the first and central.

4.3.7 Globular proteins

Simulations up to 100 ns of small proteins (PDB code: 1vii, 2cyu, 1crn, 1enh, 1pgb, 2ci2, 2a3d, 1ubq and 1ubq) at 300 K were performed with FACTS I, II ,III and IV. Under all the FACTS parametrisations (see Supplementary Material for RMSD time series) all proteins display instability, especially in case of FACTS I. Fig. 9.91 shows a significant example of what usually happens to medium-size structures with FACTS. (See Supplementary Material for further details).

4.4 Best parametrisation of FACTS

The target of this work is to find a *unique* parametrisation for FACTS (independent on the structure involved in MD simulations). The results obtained so far for each parametrisation are summed up in Tab. 4.14. The model gives best performances at low ϵ in case of unstructured peptides (wkqa, meli). In case of pgbh (which is in equilibrium between unstructured and structured conformations) and gsgs (another reversibly folding peptides) the results are more ambiguous. Gsgs and act2 features are better reproduced with $\epsilon = 2$ (act2 is experimentally 60% folded) but FACTS III, on the basis of gsgs NOEs and the pgbh perc. of unfolded structure, is also a good choice. Structured proteins are slightly more stable with higher ϵ. In conclusion FACTS III and IV perform better, with a slight preference for FACTS IV.

4.5 Test of FACTS III: protein folding of 1igd

Despite the tendency to instability shown by FACTS with respect to globular protein structure, it shows very interesting results when applied to the folding of complex structures, such as 1igd. This molecule is relatively stable with FACTS (MD simulations show RMSD with respect to X-ray structure less than 6-7 Å) under all parametrisations, except FACTS I. For these reason protein 1igd is a good candidate for from-extended MD simulations. Four runs with FACTS III at 300 K (3-4 μs) starting from a complete extended conformation of this protein were then performed. The molecule repeatedly reached structures characterised by an RMSD (with respect to

struct.	experiment/FACTS	I	II	III	IV
ion. sid.	PMF	-	-	*	-
wkqa	FRET	-	*	-	-
wkqa	EW	-	*	-	-
meli	sec. str.	*	-	-	-
meli	FET	*	-	-	-
pgbh	HCα CS, 280 K	-	-	*	*
pgbh	HCβ CS, 280 K	-	*	-	-
pgbh	HN CS, 280K	-	-	*	-
pgbh	% unf. th. dep.	*	*	-	-
pgbh	% unf. RMSD	*	*	-	-
act2	C-helix CS	-	-	*	*
act2	C-coil CS	-	-	*	*
act2	fr. hel. CS	-	-	*	*
act2	fr. hel. TS	-	-	*	*
gsgs	HCα CS	-	-	*	*
gsgs	HCβ CS	-	-	-	*
gsgs	HN CS	-	-	*	*
gsgs	NOEs	-	*	-	*
gsgs	sec. str.	-	-	*	*
Prot.	stab.	-	-	-	*
	param. set tot. score:	4*	6*	10*	11*

Table 4.14: Summary of FACTS parametrisations performances with unstructured/structured peptides.

crystal structure) under 3 Å. See Fig. 9.92 for a combined timeseries of RMSD and native contacts of one of these simulations (and see Supplementary Material for the others). The secondary structure elements involve up to over 80% the same residue involved in the native ones (see Fig. 9.92) and the topology of the protein is correctly reproduced (see the cluster analysis in the Supplementary Material), whereas the α-helix is systematically twisted with respect to the crystal structure. However, taking into account that these outcomes are attained by FACTS under CHARMM parameter19 (implying the lack of explicit treatment of the aromatics hydrogens) and, as pointed out before, realizing that FACTS needs for sure a correction to the nonpolar

contribution to the solvation energy, these results can be considered promising.

4.6 Conclusions

It is important to point out the difference in the ability of FACTS to reproduce experimental data among the parametrisations which led to the best results. FACTS behaves best using parametrisation I and II with wkqa and melittin (unstructured peptides). FACTS III and IV are optimal with act2, pgbh and small proteins (structured peptides). Pgbh and gsgs peptides, which actually belong to a twilight zone between structured and unstructured peptides, give ambiguous results (since FACTS III gives good results). Stability of globular proteins is obtained partially only with FACTS IV.

Such a behaviour suggests FACTS to be highly sensitive to the degree (quantity) of *structure* of biomolecules (rather than the size: melittin, indeed, is made of 26 residues, gsgs of 20; wkqa is made of 12 residues and pgbh of 16) and the parameters should be chosen according to the tendency to form secondary structure.

This issue could be approximately solved using an intermediate value of ϵ, such as 1.35 (leading to $\tau \simeq 0.728$). However, FACTS behaviour with more complex structures (like 1ubq, 1enh etc.) shows on the one hand that $\epsilon = 1$ (which is the best choice for unstructured structures) sets too high electrostatic interactions for complex structures, while on the other hand $\epsilon = 2$ gives better results with more complex structures and proteins, but sets too low electrostatics for unstructured peptides.

This suggests that an improvement of FACTS (at least for the present test-case) should aim at implementing an *interdependence* between nonpolar and electrostatic interactions according to some structural information rather than focus on different choices among ϵ and γ.

Fig. 9.93 shows the loss of the small hydrophobic core of 1ubq and 1enh during FACTS simulations. The degree of burial distributions of their inner atoms actually undergo a dramatic change. This outcome, together with the studies of peptides, indicates that the source of instability that we will see in Fig. 9.91 is located in this inner region of the macromolecules and becomes more relevant according to the their

globularity (or their "degree of structure"). To solve this issue a quantity able to discriminate between peptide-like structures (whose atoms are at most exposed) and globular ones (whose atoms are mostly buried inside the structure) must be developed.

Fig. 9.94 suggests that the role played by the "degree of structure" could be approximately covered by the FACTS degree of burial itself. The more an atom is buried, the less its contribution to solvation energy should be. The more an atom is exposed, the more its contribution to solvation energy should be close to the linear SASA.

Chapter 5

Error analysis of the FACTS parametrisation

Refinement of the dataset used for the parametrisation, investigations about the fitting procedure, correction to the geometric definition of symmetry and volume measure are here discussed. The target is to find a way to minimise the error in the electrostatic atomic solvation energy. As a result it is found that the model cannot be further enhanced.

There are two possible outcomes: if the result confirms the hypothesis, then you've made a measurement. Otherwise, you've made a discovery.
E. Fermi

5.1 Summary

The conclusion of Manuscript 1 is essentially that FACTS is able to reproduce experimental data in a very wide range of different conditions, such as fluorescence experiments, chemical shifts data, folding events and so forth. Although, it clearly suffers of a, let us say, *structural astigmatism*, since **one has to adapt the parametrisation to the structure under study**. In case of unstructured peptides FACTS performs better with low internal dielectrics $\epsilon = 1$. In case of globular proteins and structured peptides FACTS performs better with high internal dielectrics $\epsilon = 2$. Since it is not feasible to apply different setups of the model according to the structure under study, a correction to this issue has to be found. At first sight, it might seem that the

peptide	size	secondary structure	best int. ϵ
wkqa	12	unstructured ($\simeq$ 80%)	low
aaqa	15	α-helix ($\simeq$ 60%)	high
pgbh	16	rev. fold. β-hairpin	high/low
gsgs	20	rev. fold. 3-str. β-sheets	high/low
meli	26	unstructured ($\simeq$ 80%)	low
proteins	30-70	well structured	high

Table 5.1: Conclusions of the systematic study of FACTS parametrisation in Manuscript 1. The best electrostatic parametrisation is not affected by the size of the involved molecule. Rather it seems the degree of structure is important.

size of the biomolecule matters. But this is actually not the case. Melittin features (unstructured peptide, 26 residues) are better reproduced with low ϵ. Gsgs features (partially structured peptide, 20 residues) are better reproduced with high ϵ. FRET experiments related to wkqa (unstructured peptide, 12 residues) are very well attained (even better than with the GBMV implicit solvent model) with low ϵ. With pgbh (mostly structured peptide, 16 residues) FACTS works better with high ϵ. Therefore, the comparison between the size and the best parametrisation shows that the latter is independent on the number of residues (see Tab. 9.26).

The size does not matter. But intriguingly, the *structure* seems to matter. What are, then, the reasons for such a behaviour? Moreover, among these reasons, which are attributable to FACTS (**intrinsic** issues)? And conversely, which are the issues related to the approximation introduced in *every* CTS/GB/SASA model (**extrinsic** issues)?

5.2 Plan of the work

The problem is not trivial and thus a top down strategy could be useful for a systematic investigation. For instance, starting from the intrinsic aspect of the problem (and then switch to the extrinsic) requires less revision of theory and will allow to study more specific problems.

Now, what are the differences between FACTS and the other solvation models? The difference is concentrated in *the way FACTS determines the Born radii*. These

are calculated by means of the original volume and symmetry measures A and B, using a fitting function.

For sure errors are introduced into the parametrisation of FACTS, due to the substitution of the fdP energies with the fitted function in the atomic solvation energy calculation process. The first task should be to quantify how much (and in which way) this discrepancies between $\Delta G^{el,sol}_{fit,i}$ and $\Delta G^{el,sol}_{fdP,i}$ affects the solvation energy (and thus the dynamics) and in general if systematic errors are present in this procedure.

Actually, the assumption that A and B contain all the geometric information about solvation is itself doubtful. We will try to enhance this definition and to evaluate the effect for the electrostatic part of the solvation energy. These insights are specifically related to FACTS.

As pointed out in the introductory section of Manuscript 1, one of the weakest points in solvation modelling is the *treatment of the nonpolar contribution to solvation energy*. The last part of this work will hence be devoted to the study of an improvement of this contribution, related to a more general theory than SAS.

The top down strategy we adopted to approach the FACTS improvement can be summarised as follows:

- Checking the FACTS fitting errors (intrinsic)
- Enhancement of the definition of A and B measures (intrinsic)
- Improvement of the nonpolar interactions treatment (extrinsic)

5.3 Fitting errors: Analysis of FACTS parametrisation

In this section we perform the error analysis of the fits seen in the chapter 3. Actually, this analysis will inform us about the ability of FACTS to get the "right" solvation energy on the basis of A and B measures (the benchmark being the fdP atomic energies). Fig. 5.1- 5.7 summerise the complete error analysis of Fig. 3.5- 3.11. For each of the 7 atom types, the difference between fdP energies and the fitted values obtained using the best fit of Eq. 3.3.2 are shown as a function of A and B and as a function of fdP. The first two plots reveals the accuracy of the fit as a function of A and B (independent variables), while the third one shows the accuracy in the final determination of the fdP

energies (dependent variable).

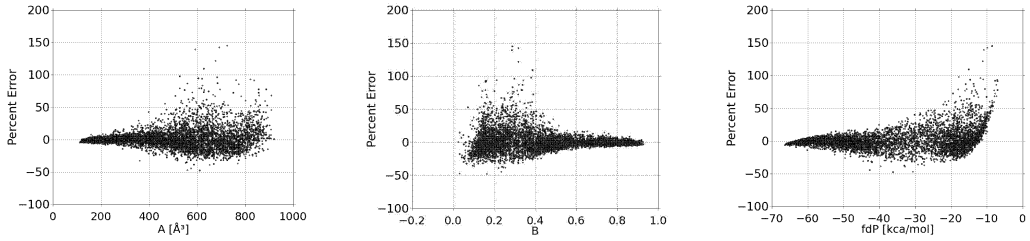


Figure 5.1: Error distribution related to the fitting in Fig. 3.5 (5827 data points, RMSE=3.529, H* atoms). (Left) Percentage error as a function of the FACTS volume measure A; (Center) percentage error as a function of the FACTS symmetry measure B; (Right) percentage error as a function of fdP, i.e. percentage error (in the atomic energy evaluation) as a function of the solvation energy.

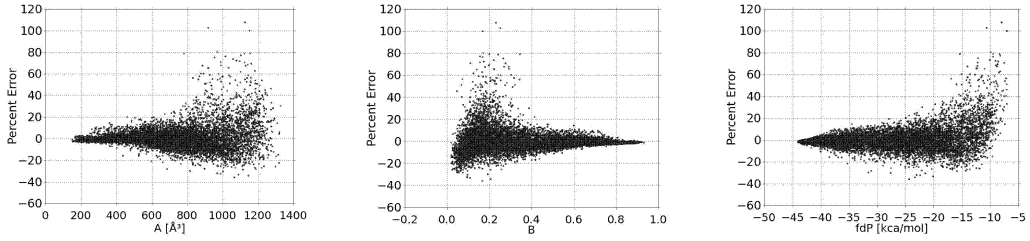


Figure 5.2: Error distribution related to the fitting in Fig. 3.6 (9164 data points, RMSE=1.873, N*, O* atoms). (Left) Percentage error as a function of the FACTS volume measure A; (Center) percentage error as a function of the FACTS symmetry measure B; (Right) percentage error as a function of fdP, i.e. percentage error (in the atomic energy evaluation) as a function of the solvation energy.

5.3.1 Conclusions of the error analysis of the original parametrisation

Fig. 5.8 and 5.9 summarise the conclusion of Fig. 3.5- 3.11, showing a common trend with all the FACTS atom types. The solvation energies are indeed systematically non sufficiently accurate for those atoms whose solvation energy is closer to zero (i.e. which are more buried inside the protein). In particular, between $\Delta G \simeq -20$ and $\Delta G \simeq -10$ the error rises exponentially from almost 0 to +60%, an overestimation that is not negligible. This means that FACTS fitting procedure gives systematically higher values of (electrostatic) solvation energy for those atoms which are more buried in the protein. This results is important because this kind of error discriminates the degree of globularity of the studied structures, regardless to its size. At the same time, the magnitude of this effect is relevant, being around 60% for the most buried atoms.

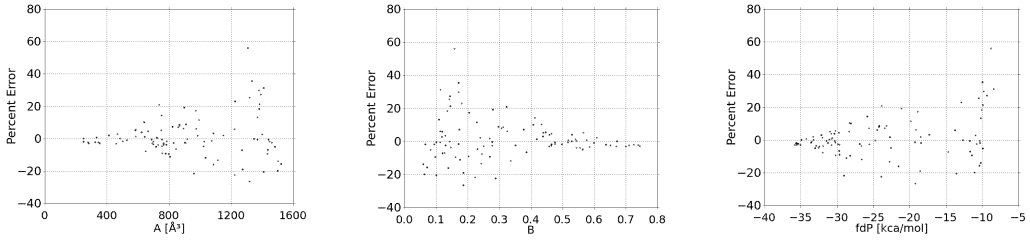


Figure 5.3: Error distribution related to the fitting in Fig. 3.7 (99 data points, RMSE=2.087, S atoms). (Left) Percentage error as a function of the FACTS volume measure A. (Center) percentage error as a function of the FACTS symmetry measure B. (Right) percentage error as a function of fdP, i.e. percentage error (in the atomic energy evaluation) as a function of the solvation energy.

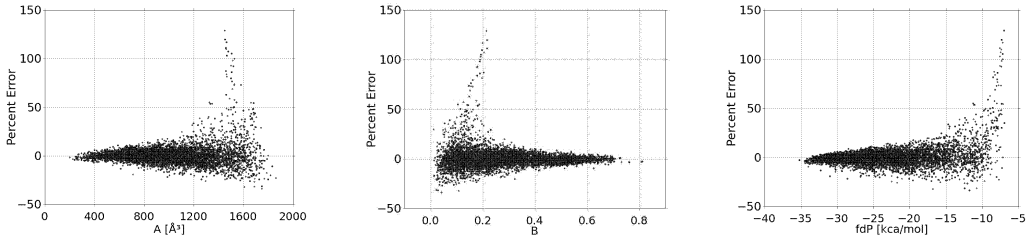


Figure 5.4: Error distribution related to the fitting in Fig. 3.8 (6264 data points, RMSE=1.399, C atoms). (Left) Percentage error as a function of the FACTS volume measure A. (Center) percentage error as a function of the FACTS symmetry measure B. (Right) percentage error as a function of fdP, i.e. percentage error (in the atomic energy evaluation) as a function of the solvation energy.

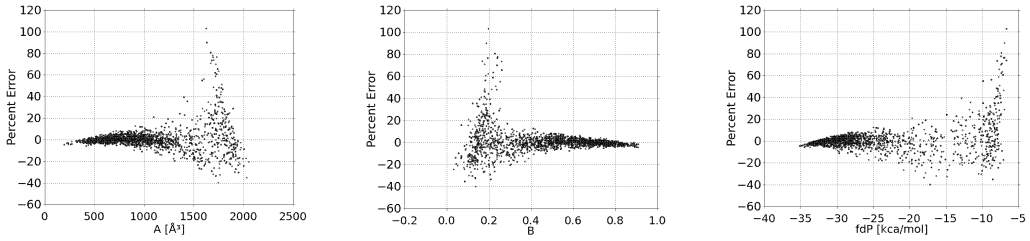


Figure 5.5: Error distribution related to the fitting in Fig. 3.9 (1750 data points, RMSE=1.375, CH3E atoms). (Left) Percentage error as a function of the FACTS volume measure A; (Center) percentage error as a function of the FACTS symmetry measure B. (Right) percentage error as a function of fdP, i.e. percentage error (in the atomic energy evaluation) as a function of the solvation energy.

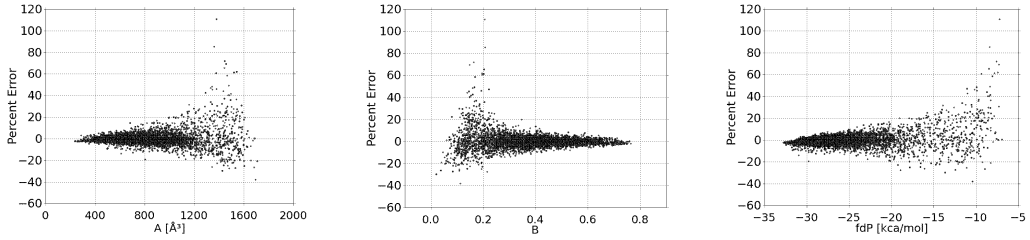


Figure 5.6: Error distribution related to the fitting in Fig. 3.10 (4118 data points, RMSE=1.164, CH2E atoms). (Left) Percentage error as a function of the FACTS volume measure A. (Center) percentage error as a function of the FACTS symmetry measure B. (Right) percentage error as a function of fdP, i.e. percentage error (in the atomic energy evaluation) as a function of the solvation energy.

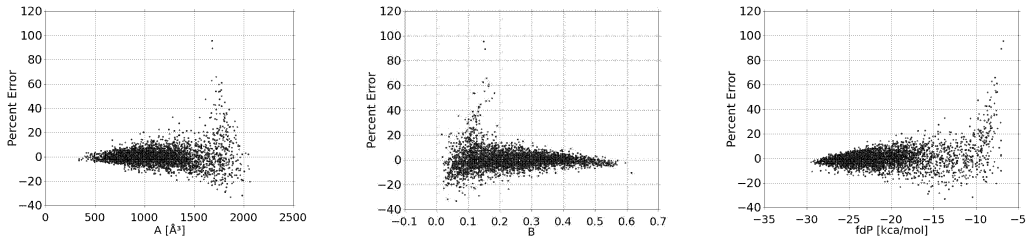


Figure 5.7: Error distribution related to the fitting in Fig. 3.11 (3814 data points, RMSE=1.160, CH1E atoms). (Left) Percentage error as a function of the FACTS volume measure A. (Center) percentage error as a function of the FACTS symmetry measure B. (Right) percentage error as a function of fdP, i.e. percentage error (in the atomic energy evaluation) as a function of the solvation energy.

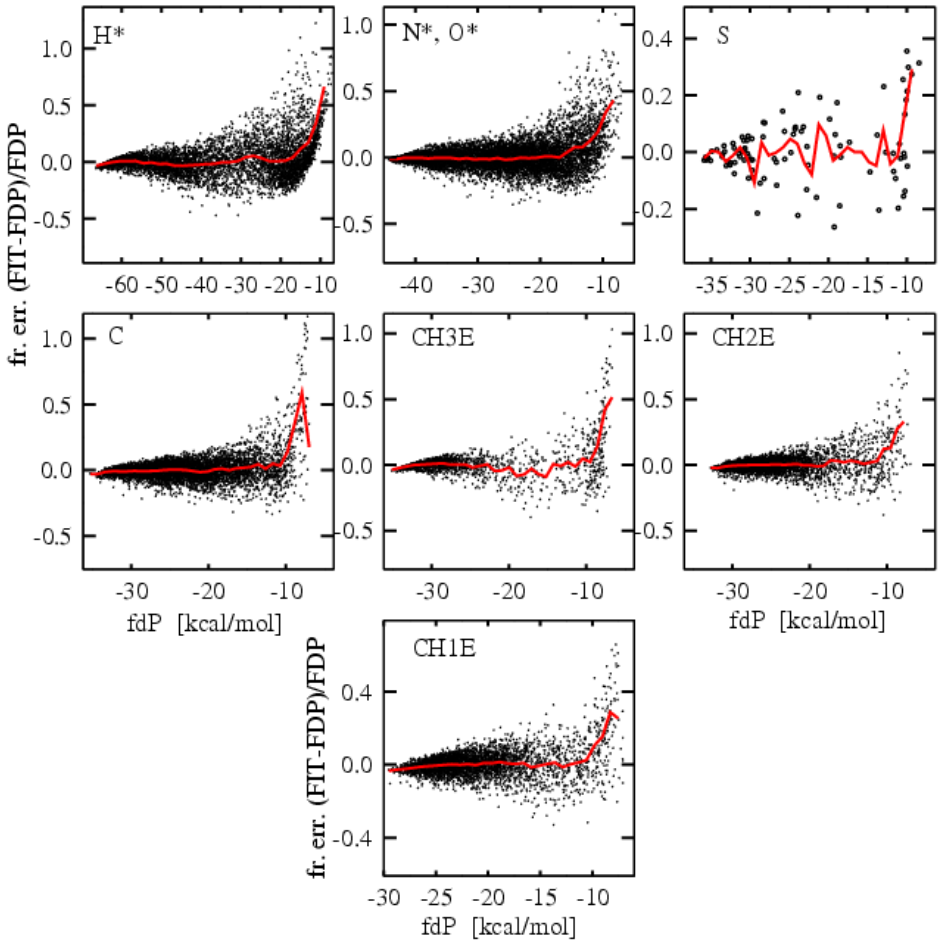


Figure 5.8: Synthesis of the error (FIT-FDP)/FDP distributions of Fig. 3.5- 3.11. For high values of fdP energies, the error in the fitting procedure rises up to 50-70%, meaning that inner atom solvation energies are less properly fitted with the selected function of A and B with respect to those of exposed atoms (the red curves represent the average value). This is confirmed by the fact that error analyses with respect to A give higher uncertainties for higher values of A (lots of atoms in the R^s sphere), while the one performed with respect to the symmetry measure B gives higher errors for atoms with $B \simeq 0 \div 0.2$ (corresponding to symmetric disposition in the R^s sphere). See also next plot (Fig. 5.9) for further details.

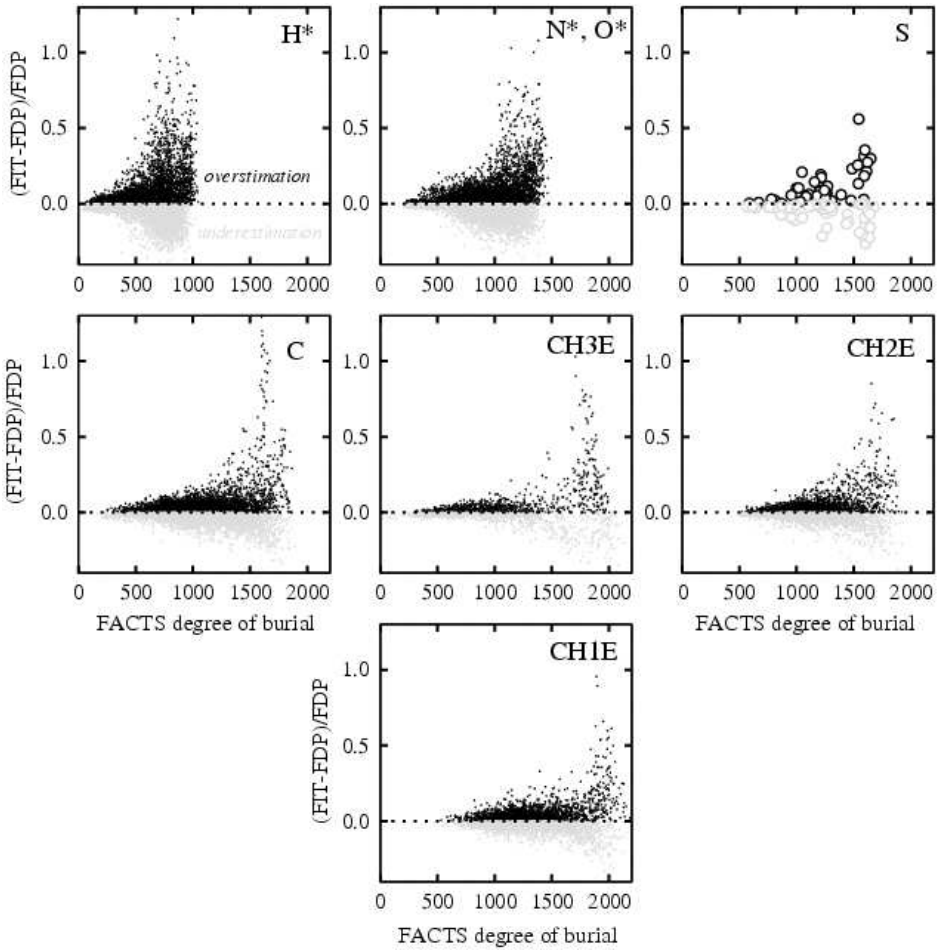


Figure 5.9: A confirmation of the previous plots, showing that the discrepancy between the fitted sigmoidal function and the fdP (atomic) solvation energies increases with the degree of burial C. Here we plot the ratio (FIT-FDP)/FDP, so that a positive value in the scale means that the fit overestimates the energy and a negative value represents an underestimation.

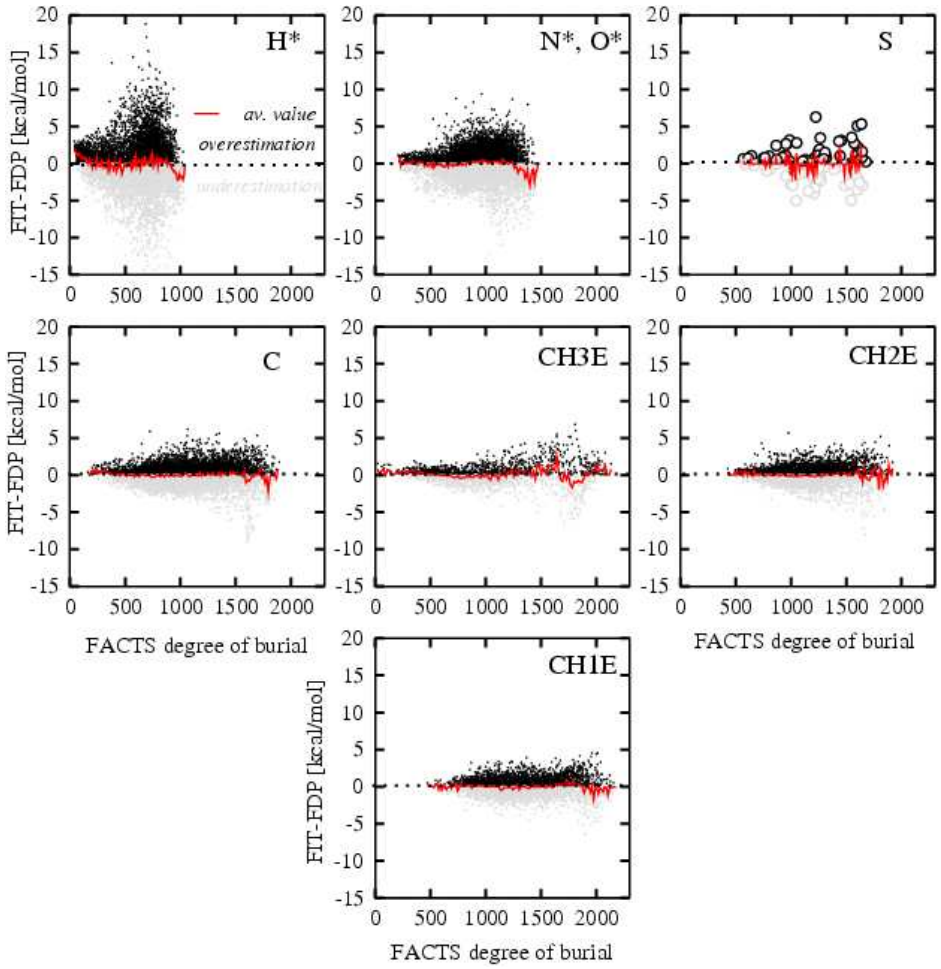


Figure 5.10: Here we plot the ratio FIT-FDP, so that a positive value in the scale means that the fit overestimates the energy and a negative value represents an underestimation.

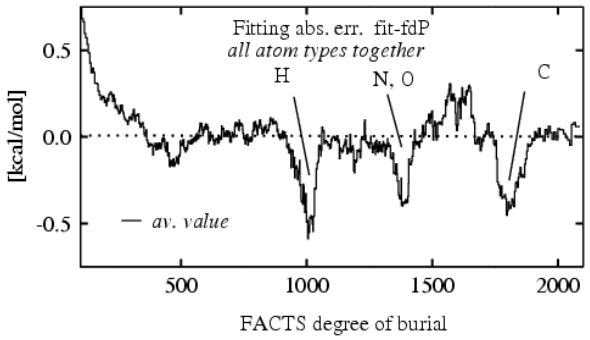


Figure 5.11: Average discrepancy between the fitted sigmoidal function and the fdP (atomic) solvation energies as a function of the degree of burial and regardless for the FACTS atom type.

5.3.2 Refining the fdP calculation

As pointed out in the previous chapter, the (A, B) space was not sampled uniformly. To better sample the (A, B) space we chose 36 very different proteins, which are 1a2p 1bpi 1crn 1dvd 1f8a 1hdn 1inc 1l2y 1pgb 1shg 1ycq 2a3d 2ins 3app 5hvp bet1 hlx1 prph 1abz 1cb3 1cus 1enh 1fmk 1hel 1kvd 1lz1 1pht 1ubq 1ycr 2ci2 2ptl 3pte anki gsgs ins2. For each of these structures we extracted 101 different conformations in such a way that the (A, B) space was more uniformly explored (see Fig. 5.12 and compare it with Fig. 3.4), in order not to weight too much the information coming from geometric redundancies[1]. This gave us a testcase of about 4000 data points for each atom type. The fdP energies used in the FACTS fitting procedure were calculated with a grid of 0.2 Å. To enhance the fit further we refined the fdP calculation up to 0.1 Å. In Fig. 5.15 we discuss the effect of the refinement (and the uniform sampling) on the fdP fitting.

5.4 Definition of A and B: The overlapping spheres problem

As one can easily argue from Fig. 3.2, the summation of volumes within the R^s sphere will result in an *overestimation* of the occupied volume, because at a big number of *intersections* between the vdW bowls. In order to evaluate the exceeding volume computed by the A_i measure, we can simply compare it to the evaluation of the *union*

[1] See for instance, Cowtan et al., *Density modification for macromolecular phase improvement*, Progress in Biophysics & Molecular Biology, 72, 245-270, (1999); Tendulkar et al., *geometric invariant-based framework for the analysis of protein conformational space*, Bioinformatics Advance Access (2005)

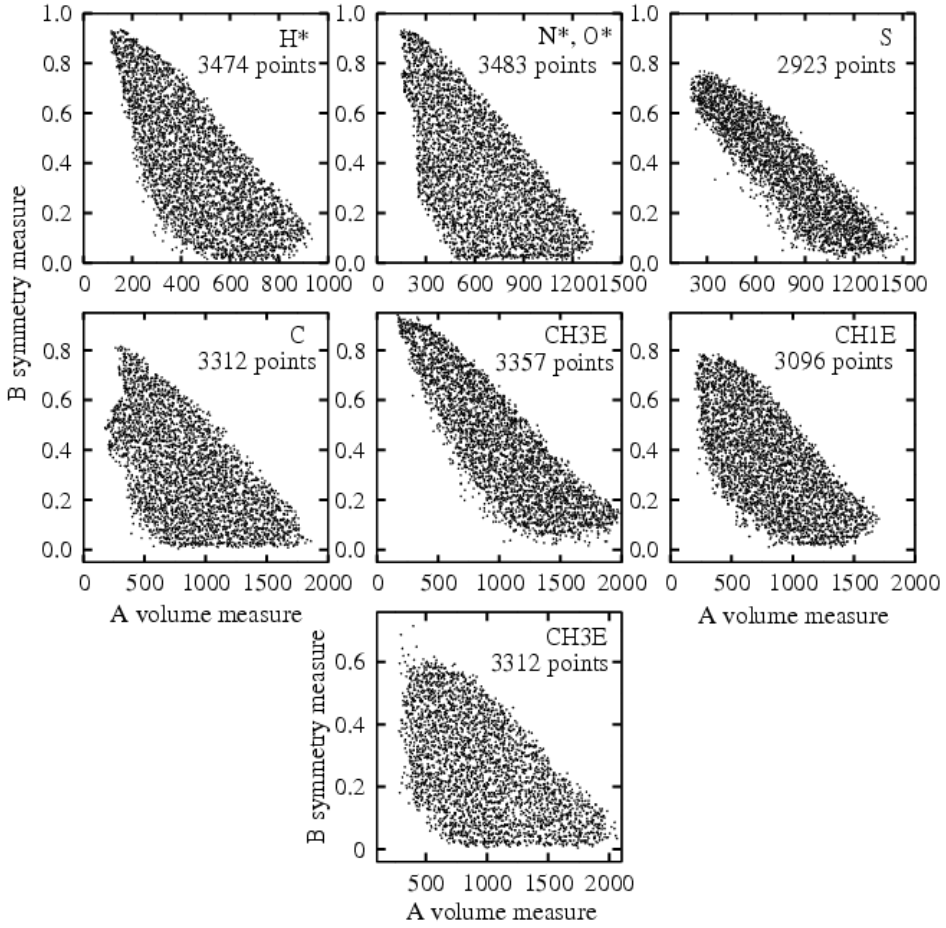


Figure 5.12: Distribution of A and B in the chosen testcase for each atom type. The structures mentioned in the text were selected in such a way that A, B explored the (A, B) plane more uniformly (within their intrinsic range of variation).

of spheres in the R^s region. By means of the coor volu command in CHARMM, it is possible to estimate this quantity. We call it UA_i. In order to take into account the presence of the Θ function, we computed:

$$UA_i = \sum_{j \in R^s}^{N} UA_j \cdot \Theta_{ij}, \tag{5.1}$$

where UA_j is the contribution to the union of volumes in R^s coming from the j atom in R^s (it is calculated by subtracting from the union of all atoms in R^s the union of all atoms in R^s but the j one). Fig. 5.13 shows the result of such a comparison. The A measure overestimates the U measure between 2.5 and 10 times.

The symmetry term B defined in Eq. 3.5 also changes if the term V_j is substituted

with UA_j. We call this term UB. Fig. 5.14 sums up the results.

Here follows the CHARMM code used to calculate the corrected UA and UB quantities.

```
SCALAR WMAIN = RADIUS
COOR COPY COMP
SCALAR WCOMP = RADIUS
SCALAR WCOMP STORE 1
SCALAR WCOMP SET 5.0
SCALAR WCOMP STORE 2
SET AJ 0
SET NUMBJX 0
SET NUMBJY 0
SET NUMBJZ 0
SET DENBJ  0
SET K 1
LABEL JLOOP          ! CYCLE OVER ALL RS-SPHERE ATOM
CALC RR = ?DIST / @R
CALC QQ = @RR * @RR
CALC MM = 1 - @QQ
CALC TT = @MM * @MM
CALC VT = @TT * @TOTMINJ
CALC AJ = @AJ + @VT
CALC NUMBJX = @NUMBJX + ( @VT * ?XAXI / ?DIST ) / ?DIST
CALC NUMBJY = @NUMBJY + ( @VT * ?YAXI / ?DIST ) / ?DIST
CALC NUMBJZ = @NUMBJZ + ( @VT * ?ZAXI / ?DIST ) / ?DIST
CALC DENBJ  = @DENBJ  + ( @VT ) / ?DIST
INCR K BY 1
IF @K .LE. @NJ THEN GOTO JLOOP
CALC NUMBJXQ = @NUMBJX * @NUMBJX
CALC NUMBJYQ = @NUMBJY * @NUMBJY
CALC NUMBJZQ = @NUMBJZ * @NUMBJZ
CALC SUMNUMJ = @NUMBJXQ + @NUMBJYQ + @NUMBJZQ
CALC SQNUMNJ = SQRT ( @SUMNUMJ )
CALC NUMDENJ = @SQNUMNJ / @DENBJ
SET FINALBJ @NUMDENJ ! UNON-SYMMETRY TERM
SET FINALAJ @AJ      ! UNION-VOLUME  TERM
```

We used then UA and UB to fit the fdP energies. In Fig. 5.15 we discuss whether these new definition are useful in order to reduce the errors between the fitted function and the fdP energies.

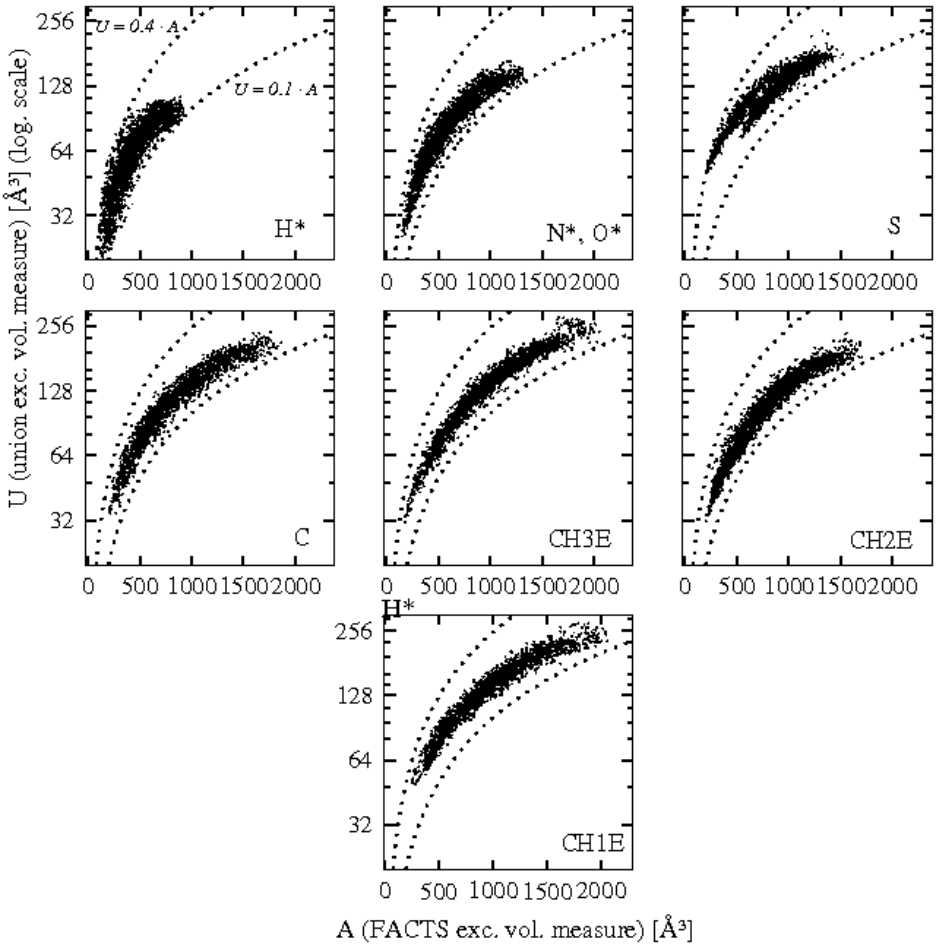


Figure 5.13: As a consequence of the spheres overlapping within the R^s region, the FACTS measure of the excluded volume A (Eq. 3.2), based on the naïve summation of vdW volumes V_j inside R^s sphere, overestimates (for all the atom types, meaning for all the R^s values) the occupied volume up to a factor of 10, with respect to the more precise estimation based on the union of vdW volumes, UA (Eq. 5.1).

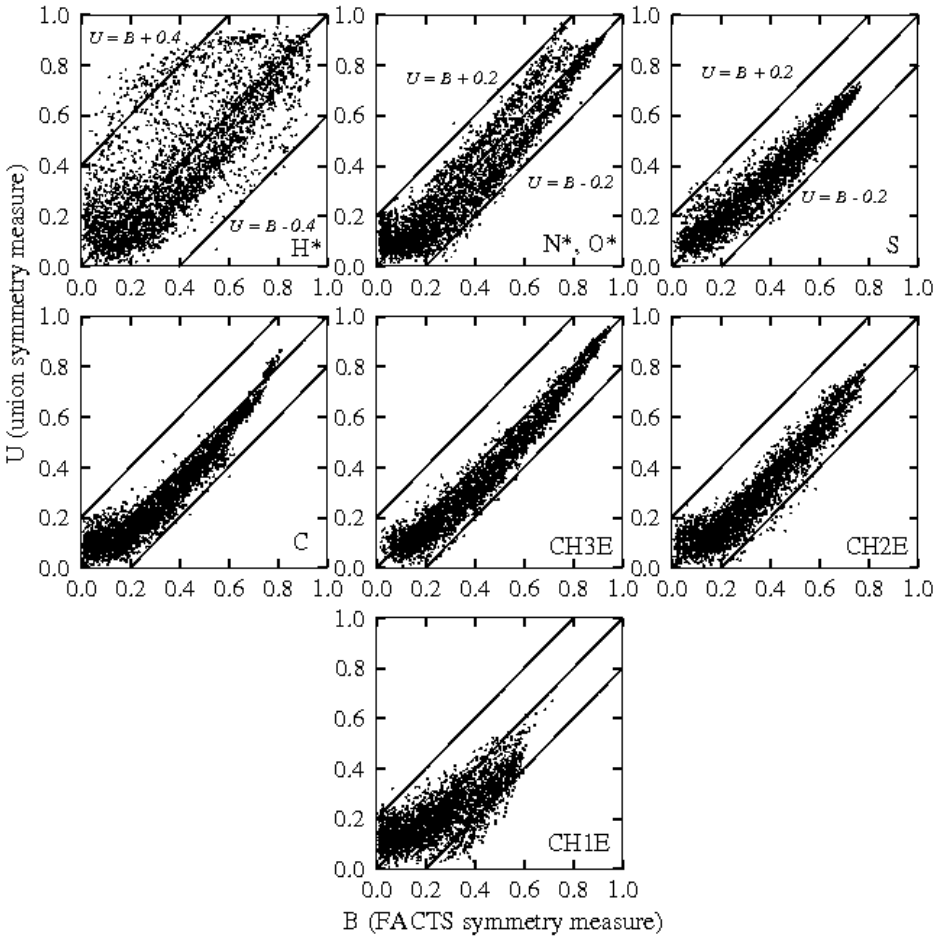


Figure 5.14: Comparison between the FACTS symmetry measure B with respect to UB, i.e. the one obtained with the substitution of UA in Eq. 3.5. It is clear that the FACTS measure of symmetry differs from the one obtained using the union of the inner volumes, especially for the hydrogens, for which the root mean square deviation between the two measures is actually of about 20%.

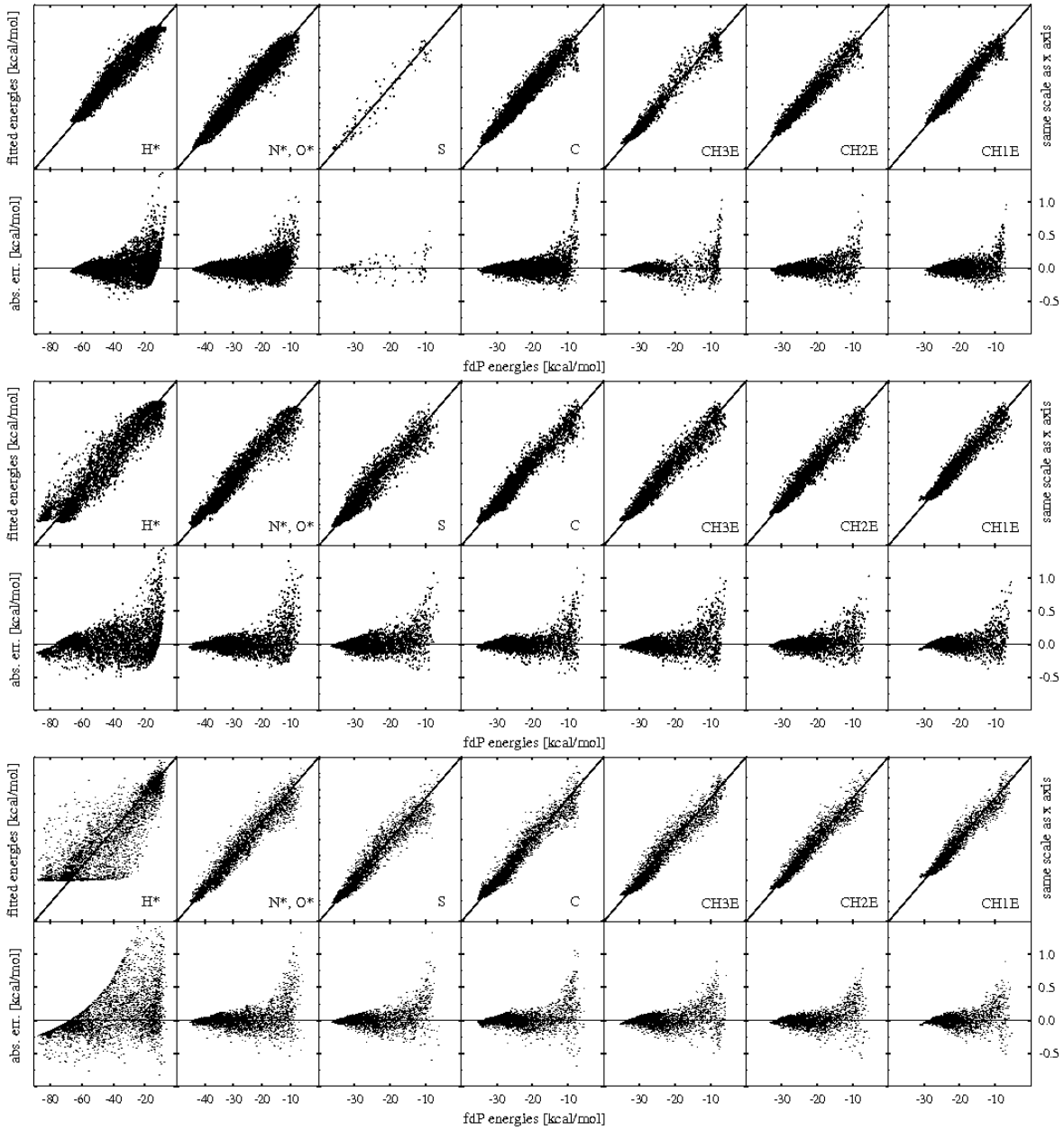


Figure 5.15: Comparison between errors in the atomic (electrostatic) energy calculations. The upper plot is related to the original version of FACTS (calculated with grid spacing equal to 2 Å). The central plot shows the error distribution in the case of uniform and refined sampling of the (A, B) space (calculated with grid space 1 Å). The bottom plot is related to the uniform and refined sample, with the geometrical correction avoiding the sphere overlapping. This comparison shows that errors in the fitting procedure (i.e, between the fitted and the calculated $fdP_{0.2\,\mathring{A}}$ energies) neither come from the sampling procedure nor from the precision of the grid spacing, nor from the geometric definition of the volume and symmetry measure. The related error distributions are actually not significantly different.

5.4.1 Best (A, B, $\mathbf{fdP}_{0.1\,\mathring{A}}$) and ($UA$, UB, $\mathbf{fdP}_{0.1\,\mathring{A}}$)

In order to perform the fit with fdP data, we have always used the original fitting function, a 3D quasi-sigmoidal function of and A and B. It is interesting to test if another functional form could better represent the relationship between A, B (or UA, UB) and fdP. We choose a large set of functional form between A, B and fdP (see TAb. 5.2). In order to avoid both overfitting and parameters number, we selected only functional forms with up to 6 parameters. The fitting procedure here is based on classical statistical tools, rather than on particle swarm optimisation. This study takes place by the needing of carefully investigating the issues in the 2007 model. A screening of 26 different functional forms (involving different logarithmic, power, rational and sigmoidal functions) allowed us to determine a best choice for the function fdp=fdp(A,B). All the calculations were performed with the online tool zunzun.com[2]. Tab. 5.3, 5.4 5.5 summarise the result of this analysis for the original, the refined and the corrected model, respectively.

5.4.2 Conclusion of the error analysis on the refined, uniform and corrected model

The conclusion of such a study is that the electrostatic setup of FACTS with parameter 19 seems not to allow for easy improvements, since geometry enhancement and fdP refinement did not reduce significantly the discrepancy between the fitted and the calculated atomic solvation energies. This insight suggests us to address the refinement of the model via the nonpolar part of solvation energy.

[2] http://zunzun.com/

code	$\Delta G_{fit,i}^{el,FACTS}(A,B)$
L.1	$a + b \cdot ln(A) + c \cdot ln(B) + d \cdot ln(A)^2 + e \cdot ln(B)^2$
L.2	$a + b \cdot ln(A) + c \cdot ln(B) + d \cdot ln(A)^2 + e \cdot ln(B)^2 + f \cdot ln(A) \cdot ln(B)$
L.3	$AB/(a + b \cdot ln(A) + c \cdot ln(B) + d \cdot ln(A)^2 + e \cdot ln(B)^2)$
L.4	$AB/(a + b \cdot ln(A) + c \cdot ln(B) + d \cdot ln(A)^2 + e \cdot ln(B)^2 + f \cdot ln(A) \cdot ln(B))$
L.5	$AB/(a + b \cdot ln(A) + c \cdot ln(B) + d \cdot ln(A)^2 + e \cdot ln(B)^2) + f$
L.6	$1/(a + b \cdot ln(A) + c \cdot ln(B) + d \cdot ln(A)^2 + e \cdot ln(B)^2 + f \cdot ln(A) \cdot ln(B))$
L.7	$1/(a + b \cdot ln(A) + c \cdot ln(B) + d \cdot ln(A)^2 + e \cdot ln(B)^2)$
L.8	$a + b \cdot ln(A) + c \cdot ln(B) + d \cdot ln(A)^2 + e \cdot ln(B)^2$
P.1	$a + A^b + B^c$
P.2	$a + A^b \cdot B^c$
R.1	$(a + bA + cB)/(1 + dA + eB)$
R.2	$(a + bA + cB)/(1 + dA + eB) + f$
R.3	$(a + b \cdot ln(A) + c \cdot ln(B))/(1 + dA + eB)$
R.4	$(a + b \cdot ln(A) + c \cdot ln(B))/(1 + dA + eB) + f$
R.5	$(a + bA + cB)/(1 + d \cdot ln(A) + e \cdot ln(B))$
R.6	$(a + bA + cB)/(1 + d \cdot ln(A) + e \cdot ln(B)) + f$
R.7	$(a + b \cdot ln(A) + c \cdot ln(B))/(1 + d \cdot ln(A) + e \cdot ln(B))$
R.8	$(a + b \cdot ln(A) + c \cdot ln(B))/(1 + d \cdot ln(A) + e \cdot ln(B)) + f$
S.1	$a + (b/(1 + e^{c \cdot (A + d + e \cdot B + f \cdot A \cdot B)}))$
S.2	$a + (b/(1 + e^{c \cdot (A \cdot d + e \cdot B + f \cdot A \cdot B)}))$
S.3	$a/((1 + e^{(b - cA)}) \cdot (1 + e^{(d - eB)}))$
S.4	$a/((1 + e^{(b - cA)}) \cdot (1 + e^{(d - eB)})) + f$
T.1	$a + bA + cB + dA^2 + eB^2 + fAB$
T.2	$a + b/A + cB + d/A^2 + eB^2 + fB/A$
T.3	$a + bA + c/B + dA^2 + e/B^2 + fA/B$
T.4	$a + b/A + c/B + d/A^2 + e/B^2 + f/(AB)$

Table 5.2: List of the 25 target function (with up to 6 parameters) used to verify that the original fitting function is the best target function. The selected functional form are classified in logarithmic (L), power (P), rational (R), sigmoidal (S) and Taylor series (T). The **S.1** code is related to the original FACTS function.

R.	H*	E.	N,O	E.	S	E.	C	E.	CH3E	E.	CH2E	E.	CH1E	E.
1	**S.1**	3.53	R.7	1.85	T.1	2.01	T.1	1.38	S.4	1.36	T.1	1.16	T.1	1.14
2	S.2	3.56	R.8	1.85	L.1	2.06	**S.1**	1.40	S.3	1.36	**S.1**	1.16	**S.1**	1.16
3	R.7	3.70	**S.1**	1.87	**S.1**	2.09	S.2	1.40	**S.1**	1.37	L.1	1.17	S.2	1.17
4	R.8	3.70	S.2	1.88	S.2	2.14	R.1	1.42	R.8	1.40	R.2	1.19	R.6	1.18
5	S.4	3.75	T.1	1.91	R.8	2.16	R.2	1.42	R.7	1.40	R.1	1.19	R.5	1.18
6	S.3	3.75	S.4	1.91	R.7	2.16	R.6	1.43	T.1	1.42	R.8	1.19	L.1	1.18
7	L.1	3.82	S.3	1.91	T.3	2.18	R.5	1.43	R.6	1.46	R.7	1.19	R.2	1.18
8	T.2	3.83	R.6	1.92	S.3	2.19	R.8	1.43	R.5	1.46	S.3	1.23	R.1	1.18
9	L.8	3.86	R.5	1.92	S.4	2.19	R.7	1.43	L.1	1.47	R.6	1.23	R.7	1.21
10	R.6	3.91	L.1	1.92	R.6	2.20	L.1	1.43	T.3	1.49	R.5	1.23	R.8	1.21

Table 5.3: Ranking (R.) of the 10 best fitting functions (chosen among the 25 of Tab. 5.2) used to test the original (A, B, $\mathrm{fdP}_{0.2\,\mathring{A}}$) FACTS data set. They are ordered according to the absolute RMSE (i.e. the average absolute error (E.), in unity of kcal/mol). Among the 25 selected functions, none gives significantly better results than the original (S.1).

R.	H*	E.	N,O	E.	S	E.	C	E.	CH3E	E.	CH2E	E.	CH1E	E.
1	**S.1**	6.72	R.8	2.32	T.1	1.95	T.1	1.57	**S.1**	1.67	T.1	1.46	T.1	1.24
2	S.2	6.84	R.7	2.32	**S.1**	1.98	**S.1**	1.58	T.1	1.67	**S.1**	1.48	**S.1**	1.25
3	R.8	7.04	**S.1**	2.33	S.4	1.99	S.2	1.62	S.2	1.74	R.2	1.53	R.5	1.33
4	R.7	7.04	S.2	2.40	S.3	1.99	R.2	1.66	S.4	1.78	R.1	1.53	R.6	1.33
5	S.3	7.11	S.4	2.42	R.8	2.00	R.1	1.66	S.3	1.78	S.2	1.54	R.2	1.33
6	S.4	7.11	S.3	2.42	R.7	2.00	R.6	1.71	L.1	1.79	R.6	1.56	R.1	1.33
7	T.2	7.27	T.1	2.44	S.2	2.02	R.5	1.71	R.8	1.81	R.5	1.56	S.3	1.36
8	R.6	7.36	R.5	2.45	R.2	2.03	L.1	1.72	R.7	1.81	L.1	1.58	S.4	1.36
9	R.5	7.36	R.6	2.45	R.1	2.03	R.8	1.74	R.2	1.81	S.4	1.59	S.2	1.37
10	T.1	7.43	L.1	2.51	R.6	2.04	R.7	1.74	R.1	1.81	S.3	1.59	L.1	1.38

Table 5.4: Ranking (R.) of the 10 best fitting functions (chosen among the 25 of Tab. 5.2) used to test the uniformly sampled (A, B, $\mathrm{fdP}_{0.1\,\mathring{A}}$) data set. They are ordered according to the absolute RMSE (i.e. the average absolute error (E.), in kcal/mol). The uniform sampling does not change the magnitude of the error in the electrostatic (atomic) solvation energy.

R.	H*	E.	N,O	E.	S	E.	C	E.	CH3E	E.	CH2E	E.	CH1E	E.
1	S.4	11.52	S.4	2.32	**S.1**	1.84	**S.1**	1.71	**S.1**	1.73	**S.1**	1.48	**S.1**	1.34
2	S.3	11.52	S.3	2.32	T.1	1.88	T.1	1.74	T.1	1.82	T.1	1.53	T.1	1.38
3	**S.1**	11.56	R.8	2.32	S.4	1.91	R.2	1.76	S.4	1.84	R.1	1.54	R.2	1.40
4	R.7	11.63	R.7	2.32	S.3	1.91	R.1	1.76	S.3	1.84	R.2	1.54	R.1	1.40
5	R.8	11.63	T.1	2.33	L.1	1.91	S.3	1.78	R.2	1.88	S.4	1.59	S.4	1.42
6	T.1	11.63	**S.1**	2.33	R.2	1.93	S.4	1.78	R.1	1.88	S.3	1.59	S.3	1.42
7	S.2	11.70	R.6	2.38	R.1	1.93	R.8	1.80	S.2	1.88	R.8	1.63	R.6	1.43
8	R.6	11.74	R.5	2.38	R.6	2.00	R.7	1.80	L.1	1.89	R.7	1.63	R.5	1.43
9	L.1	11.74	S.2	2.38	R.5	2.00	S.2	1.80	R.6	1.97	S.2	1.63	T.3	1.45
10	R.1	11.74	R.2	2.42	R.8	2.00	R.6	1.82	R.5	1.97	R.6	1.64	R.7	1.47

Table 5.5: Ranking (R.) of the 10 best fitting functions (chosen among the 25 of Tab. 5.2) used to test the uniformly sampled (UA, UB, $\mathrm{fdP}_{0.1\,\mathring{A}}$) data set. They are sorted according to the absolute RMSE (i.e. the average absolute error (E.), in kcal/mol). The comparison with Tab. 5.3, shows that: (a) S.1 fits (UA, UB, fdP) testcase as well as (A, B, fdP). (b) Even if for some atom type (C, N, O), the S.1 absolute error is reduced down to one 1 kcal/mol with respect to the original version, for all the other atom type, namely H, the error is even worst or similar.

Chapter 6

From FACTS to FACTS SISI

Here is the second part of my original contribution on the enhancement of FACTS. The setup of an empirical correction to the solvation energy, based on the error analysis in the FACTS parametrisation. The theoretical aspects and the results of the corrected model (called FACTS SISI) are here deeply investigated. This chapter is the outline of a second paper which is going to be submitted (Supplementary Material can be found in Appendx 2).

ABSTRACT: The FACTS implicit solvent modelhas been tested with several unstructured/structured peptides and small globular proteins [10]. These studies showed that optimal parameterisations are highly sensitive to the structure of the biomolecule under study. Since a more sophisticated treatment of nonpolar interactions is crucial in order to implicitly reproduce the water surrounding of biological molecules, the nonpolar contribution to solvation energy of FACTS is here enhanced by the addition of a simple sigmoidal (SISI) function of the atom degree of burial, suggested by the Tolman theory of surface tension.

6.1 Introduction

The conclusion from the systematic study of FACTS simulations shows that the model is able to reproduce experimental data (basically fluorescence and NMR experiments) in a wide range of different conditions (unstructured conformations, reversible folding, proteins stability). Moreover, the balance between accuracy and speed is the best in comparison with the most popular solvation model [10, 53]. However, the best parametrisation (i.e. the one that best approximates the experimental data) depends

on the structure in study. In particular internal dielectrics $\epsilon = 1$ in case of unstructured peptide and $\epsilon = 2$ for proteins and reversible folding peptides (see Tab. 9.26). More accurately, the mostly unstructured condition of melittin [35] is better reproduced with low ϵ (26 residues). The reversible folding feature of gsgs (a partially structured peptide [50], 20 residues) is attained with high ϵ. FRET experimental values related to wkqa (an unstructured peptide [31, 32], 12 residues) are reached (even closer than GBMV implicit solvent model and explicit water simulations) with low ϵ. The behaviour of the 1pgb terminal β-hairpin (a mostly structured peptide of 16 residues [54]) is well reproduced with $\epsilon = 2$. Therefore, the comparison between the size and the best parametrisation shows that the latter is independent on the number of residues (see Tab. 9.26). Intriguingly, the secondary structure content is clearly in relation with the best internal dielectrics setup. However, this dependence prohibits FACTS from being structure-independent and thus a correction (which takes the secondary structure of the biomolecule in study somehow into account) has to be found. In the previous chapter, it has been shown the errors in the atomic solvation energy calculation, coming from a fitting procedure over geometric properties of surrounding atoms (and the definition itself of these geometric properties) cannot be further improved upon.

mol.	size	II str.	best ϵ
wkqa	12	unstr.	low
aaqa	15	RF 1-α-helix	high
wkqa	16	RF 1-β-hairpin	high/low
gsgs	20	RF 3-β-sheet	high/low
meli	26	unstr.	low
PROT	30-70	high str.	high

Table 6.1: Conclusions of the systematic study of FACTS: best electrostatic parametrisation is not affected by the size (number of residues) of the involved molecule, while it seems to depend on its degree of structure. RF stands for reversible folding, while PROT stands for proteins whose structure involve groups of helices and β-sheets and α-helix bundles (see Supplementary Material for further details).

However, electrostatics setup is only half of the problem, since solvation models must provide also a nonpolar contribution to solvation energy. From this point of view

FACTS is a SASA model [55, 56, 57]. The systematic study shows also that different values of surface tension γ did neither overcome the protein stability nor the different behaviour of FACTS with high or low ϵ setup (low γ made model independent on SASA and high γ overstabilized proteins). On the other hand, it has be pointed out by many authors that the nonpolar treatment is a weak spots of solvation modelling [2, 3, 4, 5, 6, 7, 8, 9].

For these reasons, a correction to FACTS nonpolar solvation energy (based on the FACTS original way to determine Born radii) has been designed, implemented and tested.

6.2 Methods

6.2.1 The FACTS degree of burial

FACTS determines the Born radii via two measures dependent on the atomic geometry around each atom i. The first measure, A_i is simply the *occupied volume* within a spherical surrounding of atom i (whose radius depends on the atom type: H, N or O, C and so on). The second geometric parameter, B_i, is a measure of the *symmetry* distribution of atoms around i. A bilinear combination of A_i and B_i (called C_i) is used by FACTS to recover Born radii R_i.

6.2.2 Beyond SASA: the Tolman theory

A fruitful application of SASA model is the linear treatment of the nonpolar contribution to the free solvation energy of an atom i:

$$\Delta G_i^{np,sasa} = \Delta G_i^{cav} + \Delta_i^{vdW} = \gamma \cdot S_i, \tag{6.1}$$

where ΔG_i^{cav} is the cavitation term and Δ_i^{vdW} contains the van der Waals contribution to the nonpolar solvation energy. A natural refinement of SASA theory has been already studied by Tolman and other authors [58, 59, 60, 61, 62, 63, 64]. Let Σ be a macromolecule surface and S its related SASA. This (ideal) surface, in each point $\vec{r}$, has a peculiar *curvature* at interface with water $\sigma(\vec{r})$, which can be defined by means of the more intuitive radius of curvature $\rho(\vec{r}) = 1/\sigma(\vec{r})$. The surface tension γ can be

(a) (b) (c)

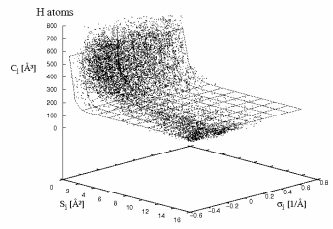
Figure 6.1: (a, b) Portion of the Σ surface of 1igd protein and (c) a simplified model, involving the relationship between S_i (i.e. the contribution to solvent accessible surface Σ, which is greater for exposed atoms), the local curvature σ_i (positive for reentrant surfaces or cavities, negative for convex surfaces, according to Tsodikov's convention) and the FACTS degree of burial C_i (the vector length in the scheme). The more an atom is buried (e.g. here atom j), the more the curvature related to the S contribution is positive, the less is S and the greater is C. Viceversa in case of exposed atoms (atom i).

written as $\gamma[\rho(\vec{r})]/\gamma[\rho(\vec{r}) \to \infty] = 1/[1 \pm a/\rho(\vec{r})] = 1/[1 + a \cdot \sigma(\vec{r})]$, where a is the radius of a water molecule ($a = 1.4$ Å for our purposes) and the sign depends on the concavity of the curvature (here, σ is negative for *convex* cavities, following Tsodikov [65]). This leads to a more refined expression of Eq. 6.1, provided the substitution $\gamma \simeq \gamma[\rho(\vec{r}) \to \infty]$:

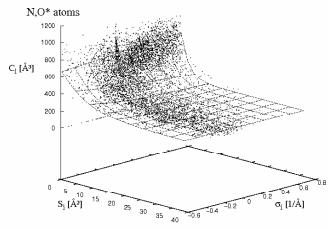
$$\Delta G_i^{np,tolm} = \frac{\gamma \cdot S_i}{1 \pm a/\rho(\vec{r}_i)} = \frac{\gamma \cdot S_i}{1 + a \cdot \sigma(\vec{r}_i)}. \tag{6.2}$$

In order to evaluate the effect of the Tolman correction, an estimation of the difference between $\Delta G_i^{np,sasa}$ and $\Delta G_i^{np,tolm}$ is useful. Indeed, from the programmer's point of view, a software implementation of such a correction into an existing model is easier this way, compared to rewriting the entire nonpolar module.

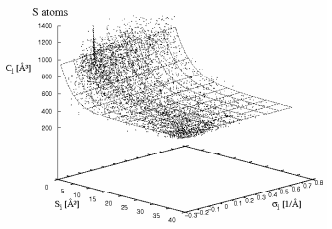
By means of the SURFACE_RACER_5.0 software tool [65], we computed the local surface curvature σ_i (i.e. the curvature of Σ_i, the contribution to Σ coming from the atom i) in correspondence of each atomic position ($\rho(\vec{r}_i) \to i$) of the following peptide-protein testcase: 1crn 1cus 1dvd 1enh 1f8a 1fmk 1hdn 1hel 1inc 1kvd 1l2y 1lz1 1pgb 1pht 1shg 1ubq 1ycq 1ycr 2a3d 2ci2 2ins 2ptl 3app 3pte 5hvp anki bet1 gsgs hlx1 ins2 prph (each in 101 different conformations). These are very different structures with widely different solvation properties [66]. In Fig. 6.1 the relationship between σ_i, S_i



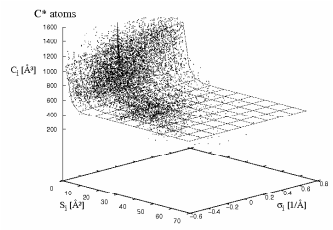
1. H* atoms



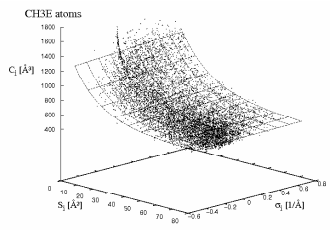
2. N,O* atoms



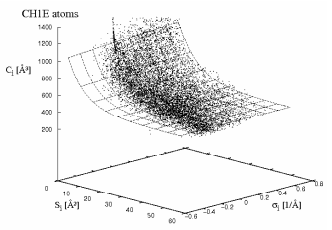
3. S atoms



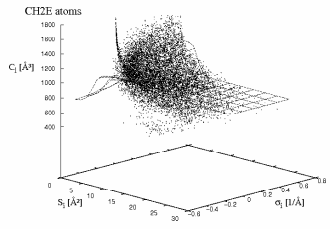
4. C atoms



5. CH3E atoms



6. CH1E atoms



7. CH2E atoms

Figure 6.2: The actual relationship between S, σ and C (surface area data are recovered by CHARMM; curvatures were calculated with SURFACE_RACER_5.0 and the degree of burial has been provided by FACTS, for each atom of the testcase described in the text). The shown 3D functions are double sigmoidal functions of S and σ fitting C.

and the FACTS degree of burial C_i (according to the atom type) is presented.

The comparison between $\Delta G_i^{np,sasa}$ and $\Delta G_i^{np,tolm}$ shows that the Tolman correction is not negligible, at least for cavities near the SAS. The difference $\Delta\Delta G_i^{np,tolm-sasa}$ ($=\Delta\Delta G_i$) between the energies varies exponentially with the surface S_i and the local curvature σ_i in a non trivial way (see Fig. 6.4). On the other hand, the relationship of

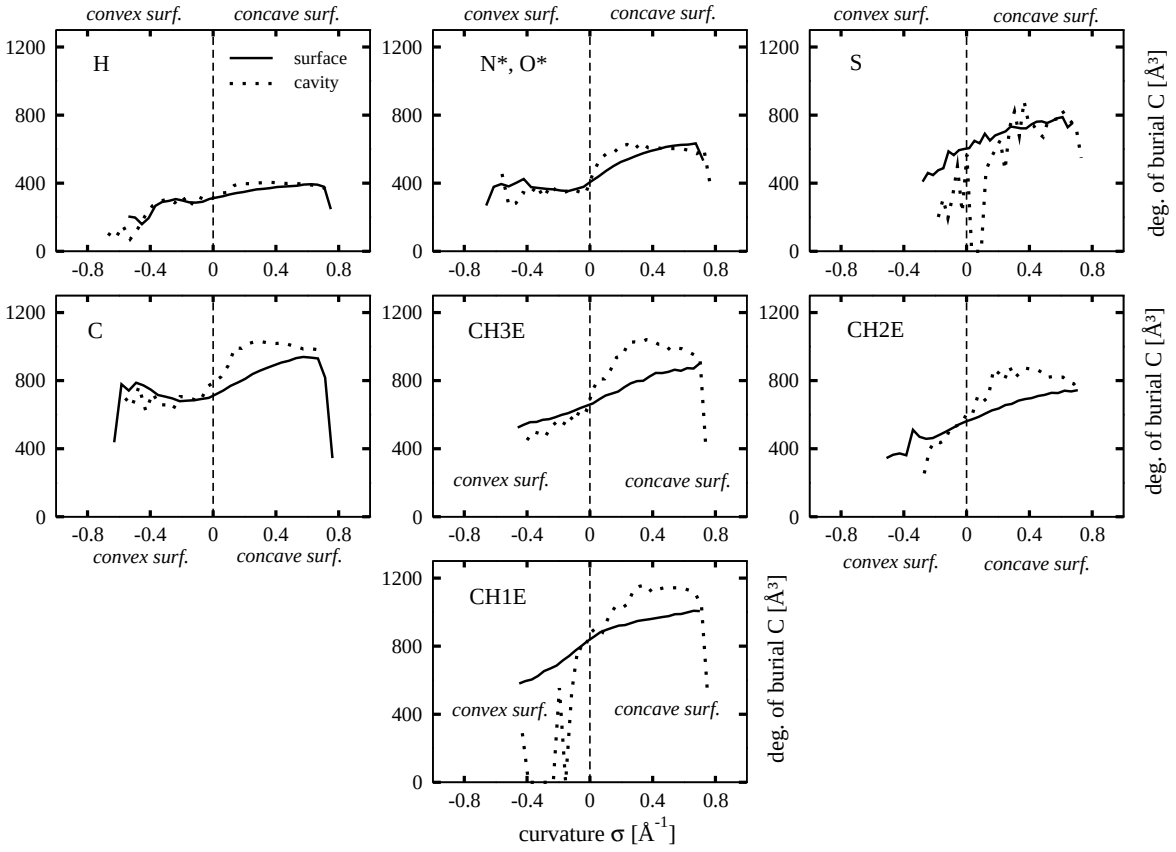


Figure 6.3: Average values of the relationship between the degree of burial C and the curvature σ seen in Fig. 6.2. Convex surfaces (negative curvature, according to Tsodikov's convention) are related to exposed atoms (see Fig. 6.1), while concave surfaces (positive curvature) are related to more buried atoms. SURFACE_RACER_5.0 can select atoms belonging to cavities. Dotted lines are related to atoms belonging to cavities which can be accessible to water. In general, the Tolman effect (dependence of the surface tension on the curvature) is then more relevant for atoms belonging to internal cavities.

such a discrepancy and the FACTS degree of burial – which can also be seen as additive correction to SASA energy – shows a remarkably clear trend (see Fig. 6.5), allowing us to setup an *averaged* (unique) correction function for all the atom types. The new model, FACTS SISI, has been tested via MD simulations over the same peptide-protein testcase of the FACTS systematic study. The computer simulations were performed by the CHARMM program (version c35a2) with the polar hydrogen parameter set PARAM19 [13] and using either a Berendsen's bath (coupling constant: 5 ps) or a

Langevin dynamics (friction constant: 0.15 ps^{-1}), always with a time step of 2 fs. The van der Waals radius of all hydrogen atoms was set to 1 Å. The CHARMM default truncation scheme of long-range electrostatics and van der Waals energy was used (SHIFT to 0 energy at 7.5 Å). The non-bonding interactions were updated heuristically. The SHAKE algorithm was used. Coordinate frames were saved every 20 ps.

6.3 Results and Discussion

The SISI model of nonpolar interactions $\Delta G_i^{SISI} = \gamma \cdot S_i + g_0 + \frac{g_1}{1+e^{-g_2 \cdot (C_i - g_3)}}$ has been implemented in FACTS. The parameters g_0, g_1, g_2 and g_3 had to be setup, with the help of the fitting function in Fig. 6.5. The MD simulations analysis shows that such a simple sigmoidal term can concurrently yield stability of medium-size proteins and solvation properties of peptides.

To test the SISI nonpolar contribution of FACTS, the following parameters were kept constant in all simulations. The internal dielectric ϵ has been set to 2 and the surface tension γ has been set to 0.0075 [kcal/(mol $\cdot$ Å^2)]. The SISI parameters are $g_0 = 0$ [kcal/mol], $g_1 = -0.7$ [kcal/mol], $g_2 = 0.02$ [Å^{-3}] and $g_3 = 1400$ [Å^3] (we will omit these units in the following). The robustness of the SISI correction was then investigated by means of four different parameter sets for the sigmoidal function (see Supplementary Material).

The target of this work is to show that the SISI correction is able to overcome the problem of a unique parametrisation for FACTS. This means in practice that it is necessary to find a compromise between the low internal dielectric setup, which works better with unstructured peptides, and the high internal dielectric setup, which is able to partially attain the stability of structured peptides and small proteins. The effects of the tool used to find such a compromise (the designed SISI correction) are compared with FACTS results related to $\epsilon = 1$ (optimal for unstructured peptides) and to $\epsilon = 2$ (optimal for protein stability). The surface tension γ is always set to 0.0075 (both for FACTS and FACTS SISI simulations) if not otherwise specified.

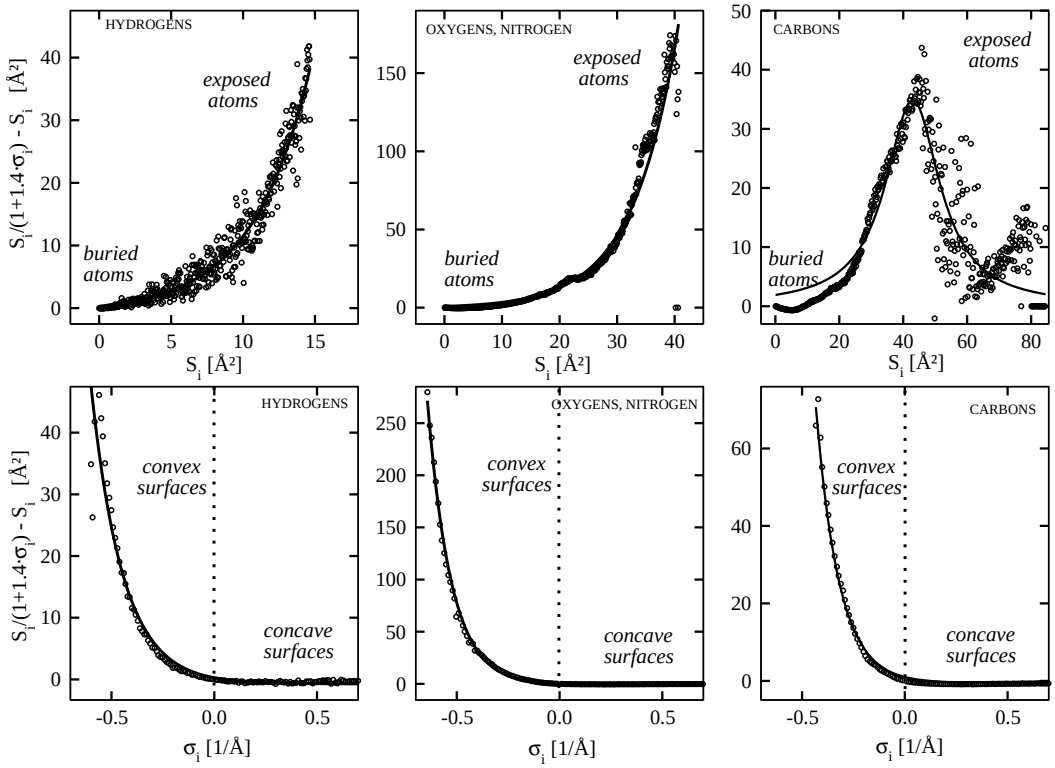


Figure 6.4: Difference between the Tolman and the SASA term for H atoms (first column), N and O (second column) and C atoms (third column) with $\gamma = 1$ kcal/(mol $\cdot$ Å^2), calculated on the basis of the S_i recovered by CHARMM all over the protein test cases in relation to the (local) curvatures σ_i calculated with SURFACE_RACER_5.0 as a function of S_i (up) and σ_i (down). For hydrogens, oxygens and nitrogens the difference is almost 0 in case of low S_i, associated with atoms far from the SAS, and in case of positive curvature, associated with reentrant surfaces (and, thus, relatively buried atoms). Viceversa, the discrepancy is high in presence of high S_i and negative curvature (the behaviour of C atoms is slightly different, since they shows a maximum error for intermediate values of S_i). Remarkably, the trend in both variables seems thus related to the degree of burial and the difference is not negligible, since, for instance, $\Delta\Delta G_{i,H}(S_i) \simeq 1.28^{S_i} - 0.69$, and $\Delta\Delta G_{i,H}(\sigma_i) \simeq 0.016^{\sigma_i} - 0.7$. The functional dependence on S_i or σ_i is different for different atom types.

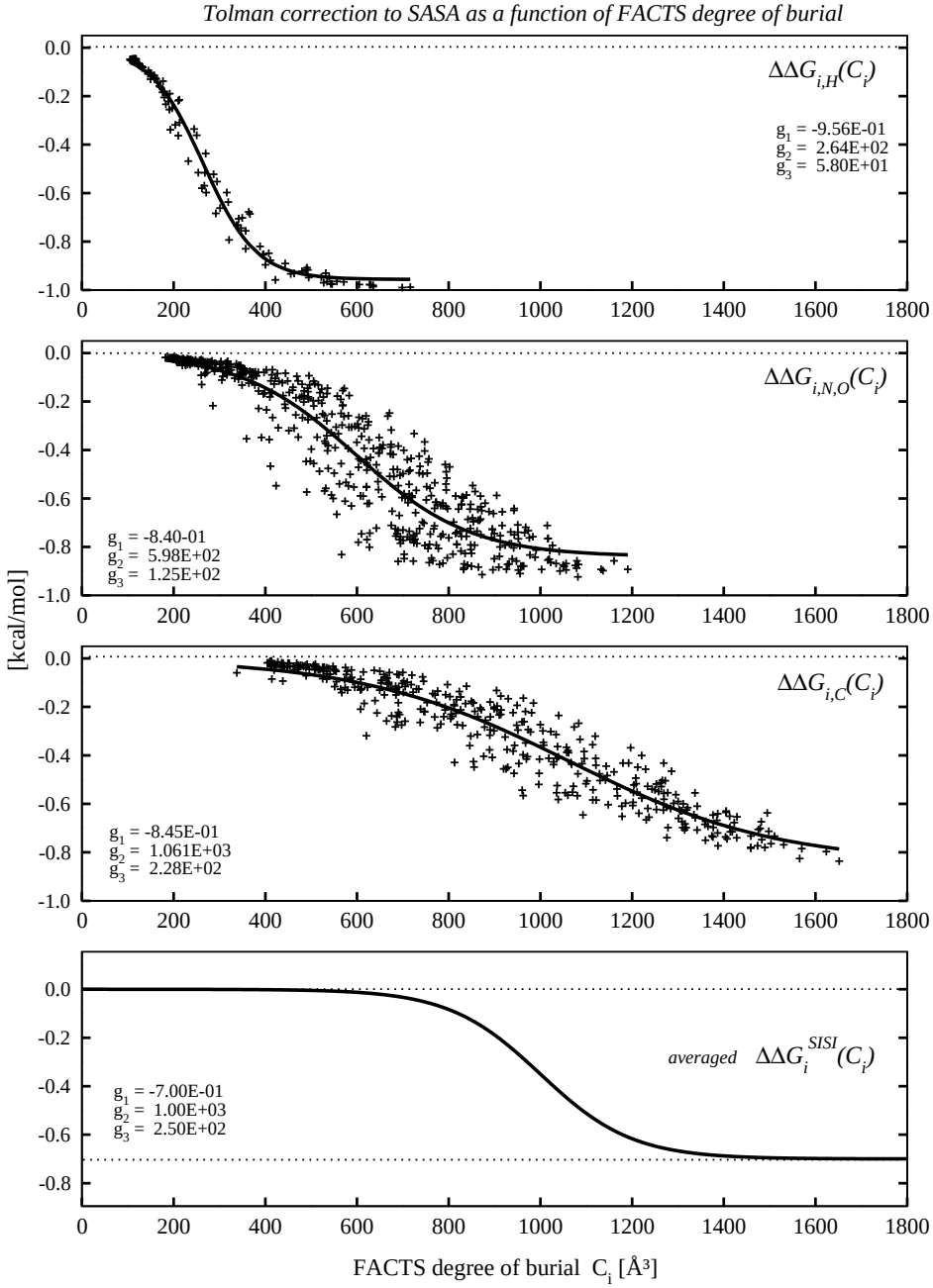


Figure 6.5: Deviation of the Tolman nonpolar (atomic) energy from the SASA nonpolar (atomic) solvation energy $\Delta\Delta$ as a function of FACTS degree of burial C for three different atom types (this is actually the correction needed to be added to SASA term in Eq. 6.1). Since the trend for all the 7 atom types involved in the CHARMM19 version is the same (this is not true when the independent variable is S_i or σ_i, see Fig. 6.4), it is meaningful to average them. The fit is made with a *simple sigmoidal* function (SISI) so that $\Delta\Delta G_i(C_i) = g_1/[1 + e^{-g_2\cdot(C_i - g_3)}]$. The averaging procedure allow us to reduce the number of free parameters.

6.3.1 Unstructured conformation of melittin

High resolution ^{1}H-NMR studies of monomeric melittin (GIGAVLKVLTTGLPAL-ISWIKRKRQQ) in aqueous solution (at 360 MHz, pH 3.0 and 30°) showed that such a polypeptide is predominantly in an unstructured and flexible form [35, 36], mostly unstructured [37] and with a low helix content [38]. Four simulations of 6 μs each were performed with FACTS starting from an extended and equilibrated structure with different internal dielectrics/surface tension parameters. The secondary structure analysis of the conformations is presented in Tab. 6.2. FACTS with $\epsilon = 2$ favours the β-strand formation between residue 1-10 and 11-24 up to 53% of the entire trajectory and favours the α-helix formation between residue 13-26 up to 22%. This is in contrast with the hypothesis of unstructured peptide. Nevertheless, the simulations related to FACTS with $\epsilon = 1$ are closer to the random-coil hypothesis [37].

secondary structure	EW	ORD	FACTS $\epsilon = 1$	FACTS $\epsilon = 2$	FACTS SISI
β strand	-	**0.00**	0.02	0.11	0.08
3-10 helix	*0.02*	-	0.06	0.11	0.10
bend	*0.15*	-	0.00	0.02	0.01
turn	*0.09*	-	0.03	0.10	0.07
random-coil	*0.29*	**0.88**	0.88	0.39	0.60
α helix	*0.42*	**0.12**	0.01	0.22	0.11
π helix	*0.03*	-	0.00	0.05	0.03

Table 6.2: Secondary structure analysis of melittin with FACTS and FACTS SISI performed with DSSP [39] program. Comparison with EW simulations [40] (italic) and optical rotary dispersion (ORD) [41]. Notice the decreasing random-coil percentage as a consequence of increasing internal dielectric and/or surface tension. π-helix, short β-bridges, turns, bends and 3-10 helix, are highly ephemeral along the trajectories and not specific to any residue. FACTS with $\epsilon = 1$ resulted as the best parametrisation choice with respect to the ability of the model to reproduce melittin random-coil percentage in water.

6.3.2 Energy landscape of end-to-end distance of wkqa

To assess the ability of FACTS and FACTS SISI to reproduce the fluorescent-resonance-energy-transfer experimental data (FRET) of the wkqa end-to-end distance (r in the following), 10 μs long simulations (Berendsen's bath) were made with FACTS using different values of internal dielectric and surface tension and then compared with FACTS SISI simulations at 300 K performed for each parametrisation. The potential of mean force (PMF) for folding the 12-residue peptide from the fully extended state to a compact state (where the peptide extremities are in close proximity) was then calculated for the FACTS and the FACTS SISI solvation model. $PMF(r) = -k_B \cdot T \cdot \ln\ f(r)$, where $f(r)$ is the relative frequency the end-to-end distance r along the trajectory. Eventually, the FRET efficiency was computed from PMFs and compared to the experimental result [31]. With a Förster critical distance R_0 equal to 23.6 Å and $E(r) = R_0^6/(R_0^6 + r^6)$ being the statistical mechanical expression for the FRET efficiency for peptides E, it can be shown that the latter is related to PMF through the expression $E \simeq \int_m^M E(r)e^{PMF(r)/kT}dr/\int_m^M e^{PMF(r)/kT}dr$, where m and M are the minimum and maximum value of r. In Tab. 6.3 the results of FRET calculations related to the FACTS and FACTS SISI are listed.

6.3.3 Helicity of acetyl-(AAQAA)$_3$-amide

The behaviour of FACTS SISI with the aaqa peptide has been studied. Simulations of two μs with Langevin dynamics at 274 K with FACTS from an extended structure confirmed better results with $\epsilon = 2$, $\gamma = 0.0075$. Analogous FACTS SISI simulations are expected to yield to the same behaviour, as the correction should not strongly affect this structure (the structure, indeed, allows only low degree of burial). Experimental data about the fraction of helicity along the peptide chain provided by Sholongo [45] are recovered via chemical shifts measurements by $f = (\delta - \delta_c)/(\delta_\alpha - \delta_c)$, δ, δ_α and δ_c being the observed (carbonyl-carbon) shift, the chemical shift of the helix conformation and the chemical shift of the coil conformation, respectively. Errors on the experimental helicity are calculated (using the gaussian error propagation) from the chemical shifts standard deviation (which is $\simeq 0.07$ ppm) using the previous expression for helicity and evaluated around 0.1 units. Fig. 6.6 shows that helicity content

source	FRET
EXP	**0.46**
EW	0.50
ACE	0.99
EEF1	1.00
GBMV	0.97
SASA	1.00
FACTS $\epsilon = 1$	0.52
FACTS $\epsilon = 2$	0.30
FACTS SISI	0.47

Table 6.3: FRET calculations related to FACTS without nonpolar correction for different values of internal dielectric/surface tension and FACTS SISI obtained using PMFs (for FACTS and FACTS SISI errors are in the last cipher). FACTS works better with low ϵ. FACTS SISI (which is setup with $\epsilon = 2$ and $\gamma = 0.0075$ but contains the SISI correction discussed above) approximates the experimental FRET even better than explicit water (EW) simulations, showing that such a sigmoidal correction based on the degree of burial and the Tolman theory plays a crucial role, with respect to the simple SASA theory implemented in FACTS.

of acetyl-$(AAQAA)_3$-amide peptide per residue, calculated with SHIFTX program [24] directly from trajectory, is close to Sholongo's data, expecially in the C-terminal region of the peptide. Experimental data suggest that the helical conformation of such a peptide is not stabilised by electrostatic interactions [45]. Therefore, the match with chemical shift data should mainly due to the nonpolar contribution to solvation energy, namely to the linear part of it, since the degree of burial of atoms belonging to act2 is too low to allow the sigmoidal correction affecting the dynamics. A comparison between FACTS SISI and SASA results shows a better agreement with Sholongo's data, meaning that FACTS model should be less helix-stabilising than SASA [67].

6.3.4 Reversible folding of a β-hairpin

The FACTS SISI behaviour with a simple β-hairpin was investigated. The fragment is derived from the B1 domain of the streptococcal energy protein, and it contains the only natural sequence which folds in a native-like β-hairpin structure in water [69, 70]. Chemical shifts were then used to verify FACTS SISI behaviour in comparison with

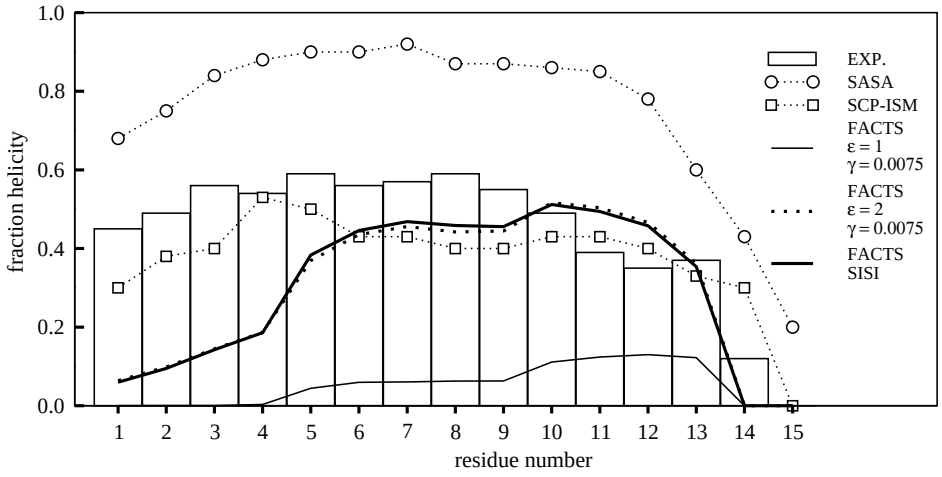


Figure 6.6: Comparison between aaqa helicity obtained by Sholongo from two-state analysis of the thermal dependence of carbonyl-carbon chemical shift measurements in pure water at 1 C° (bars) and the helicity content per residue related to SASA [49] (circles) which overstabilises helices, SCP-ISM [68] (squares) which shows the best agreement with Sholongo's data, FACTS (thin and dotted lines) and FACTS SISI (bold line). Here the influence of the SISI nonpolar correction is less strong than for wkqa; indeed, FACTS with high ϵ and FACTS SISI show a similar trend. Nevertheless, residues helicity between 4 and 10 (underestimated by FACTS) is a little higher with FACTS SISI; residue helicity between 10 and 14 (overestimated by FACTS) is a little lower with FACTS SISI.

chemical shift data [42]. The δHC_α and δHN peaks are related to each residue recovered out of four 5-μs trajectories each at 278 K.

The barycentre of each distribution has then been computed in order to compare these peaks with experimental signals (the width of each distribution gave a hint about the uncertainty of these values). The comparison has been made with Blanco's data at 278 K [42]. All the internal shifts – which are the more meaningful in this kind of analysis – are comparable with experimental data. The best results were obtained with a FACTS setup with low internal dielectrics. FACTS SISI is able to reproduce experimental resonances as accurately as FACTS with low internal dielectrics (see Fig. 6.7).

6.3.5 Reversible folding of gsgs

A three-stranded antiparallel β-sheet peptide [50] made of by 20 amino acids with experimental folding rate of μs, was used to stress FACTS and FACTS SISI towards reversible folding. Timeseries of RMSD and contacts related to 5 μs simulations are shown in Fig. 9.89. As already pointed out, this peptide should be critically affected by the nonpolar correction. Actually, FACTS SISI results in more stabilising, allowing reversible folding with a folding rate of $\simeq$ 500 ns, in contrast with FACTS ($\simeq$ 100 ns). Moreover, the comparison of NOEs violations (at 300 K) for FACTS and FACTS SISI confirms that the nonpolar SISI correction significantly improves FACTS performances (see Tab. 6.4).

model	vw	w	m	m-s	s
FACTS $\epsilon = 1$	3	6	6	2	0
FACTS $\epsilon = 2$	0	2	4	2	0
FACTS SISI	1	0	3	0	0

Table 6.4: Violation of the medium- and long-range NOE connectivities of the gsgs peptide, related to FACTS with different parametrisations and FACTS SISI. Experimental data is related to 1 mM of gsgs peptide in aqueous solution, pH 3.4, at 10 C°. See Table 1 of [50] for details. FACTS with $\epsilon = 2$ and $\gamma = 0.0075$ attained the lowest number of violations (vw=very weak; w=weak; m=medium; m-s=medium-strong; s=strong). FACTS SISI shows violations not stronger than medium, suggesting that the nonpolar correction is actually affecting the dynamics in the right direction.

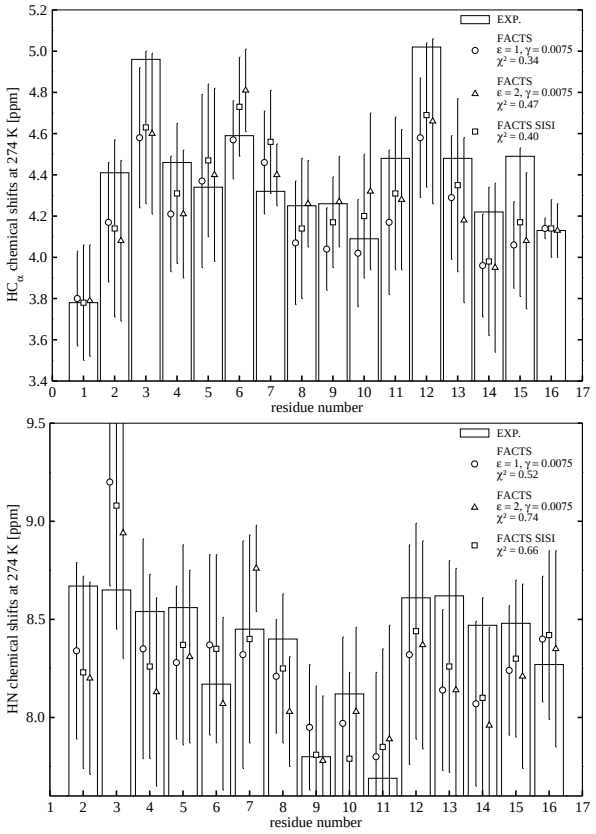


Figure 6.7: Comparison between FACTS behaviour with chemical shifts related to pgbh at 278 K (left, δHC_α peaks; right HN peaks) as reported in Ref. [42]. With FACTS, the best result is obtained with low internal dielectrics rather than with high. Nevertheless the SISI model partially corrects the FACTS trend (with $\epsilon = 2$), giving an error which is comparable to that of FACTS with low internal dielectric.

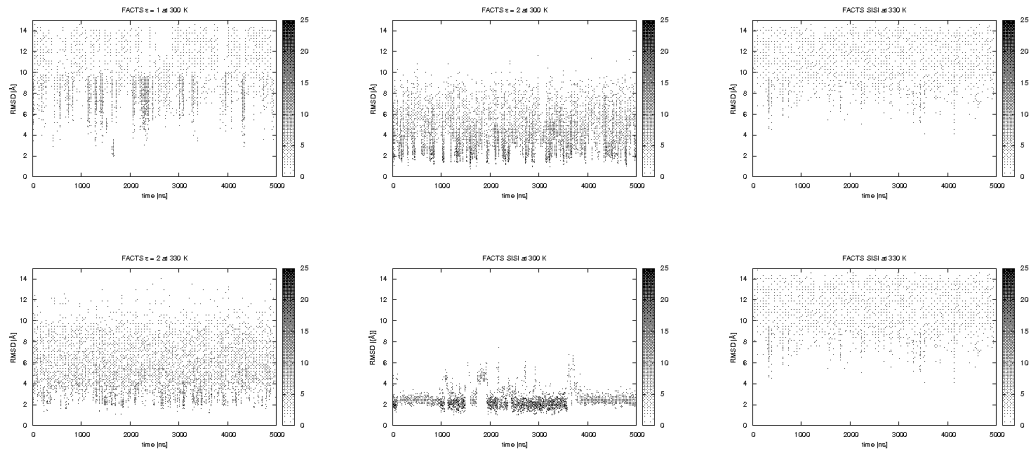


Figure 6.8: Timeseries of RMSD and contacts (black/white) of the gsgs peptide at 300 with FACTS (left, $\epsilon = 1$, center $\epsilon = 2$) and FACTS SISI (right). The comparison shows that FACTS SISI behaves better, since it leads to about 1-2 folding events within 2 μs (in this picture, a folding event occurs when a low, light-orange region is inserted between two high, dark-blue ones, or vice versa), the experimental upper limit being (at 10 C°) about 1 folding event within 5 μs [50].

6.3.6 Stability and fluctuations of small proteins

FACTS simulations (without nonpolar correction) of small proteins (2a3d, 1ubq, 1igd, 1enh, 1pht, 1vii and 2ci2) performed with all the combination between high/low internal dielectrics and high/low surface tension were not stable within 100 ns (see [53]). Fig. 9.91 displays a significant sample of what usually happens with FACTS in relation with medium-size structures (here RMSD time series of 1ubq and 1enh have been selected). Fig. 6.10 shows the *degree of burial spectra* of these simulations: it is an useful tool to investigate in which part of the protein structure the instability rises.

The RMSD time series related to FACTS SISI simulations show that the correction gets rid of the instability problem of FACTS (see Tab. 6.5). See in particular 1ubq and 1enh entries to make a comparison with FACTS. Besides, in order to verify that the SISI correction does not lead to an over stabilisation of the proteins, RMSD fluctuations, averaged every 10 ns, have been calculated from MD simulations and then compared with experimental β-factors related to C_α carbons in the PDB file, according to the relation $\beta \simeq \frac{8}{3}\pi^2 \cdot \mathrm{RMSF}^2$ (see Fig. 6.11). This study allow us to assess that protein

stability achieved by FACTS SISI is not due to overstabilization.

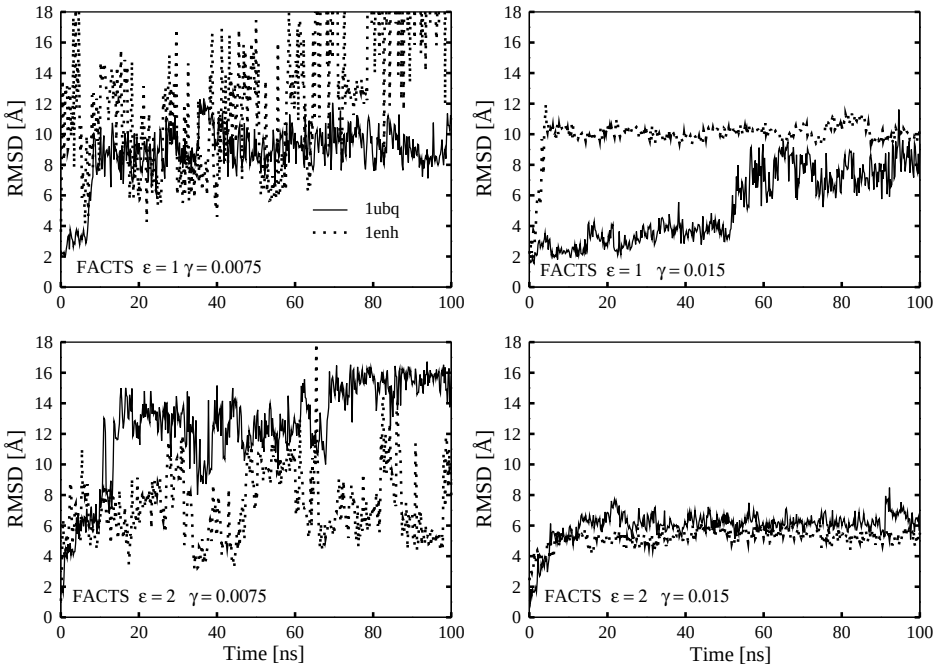


Figure 6.9: Time series of RMSD of 1ubq and 1enh with different FACTS setup (100-ns simulations of CHARMM Langevin dynamics at 300 K). The best result is obtained with $\epsilon = 2$, $\gamma = 15$, leading to a RMSD of $\simeq 9$ Å for 1ubq and 7 Å for 1enh.

p./ns	10	20	30	40	50	60	70	80	90	100
1igd	2.9	2.2	2.7	2.6	3.3	3.0	2.9	3.0	2.7	2.9
1vii	4.8	4.2	5.4	4.2	4.3	4.8	4.5	4.5	4.2	4.3
1crn	2.2	2.3	2.0	2.9	2.5	1.7	2.4	2.6	2.4	2.2
1enh	3.0	3.0	2.6	2.6	2.7	2.7	3.0	2.8	3.0	2.4
2ci2	1.6	1.8	1.7	1.5	1.9	1.6	1.7	1.7	1.8	1.4
2a3d	3.1	2.7	3.0	2.9	3.0	3.1	2.9	2.9	3.0	3.2
1ubq	2.6	1.6	1.4	1.5	1.5	2.9	3.0	3.3	3.1	3.2
1pht	2.0	2.3	1.9	2.1	2.2	2.4	1.9	2.1	1.9	1.9

Table 6.5: RMSD timeseries of the protein test-case with FACTS SISI at 300 K (Berendsen's bath). FACTS SISI obtains protein stability with the same parameter set used to obtain gsgs reversible folding and 60% of melittin unstructured conformation.

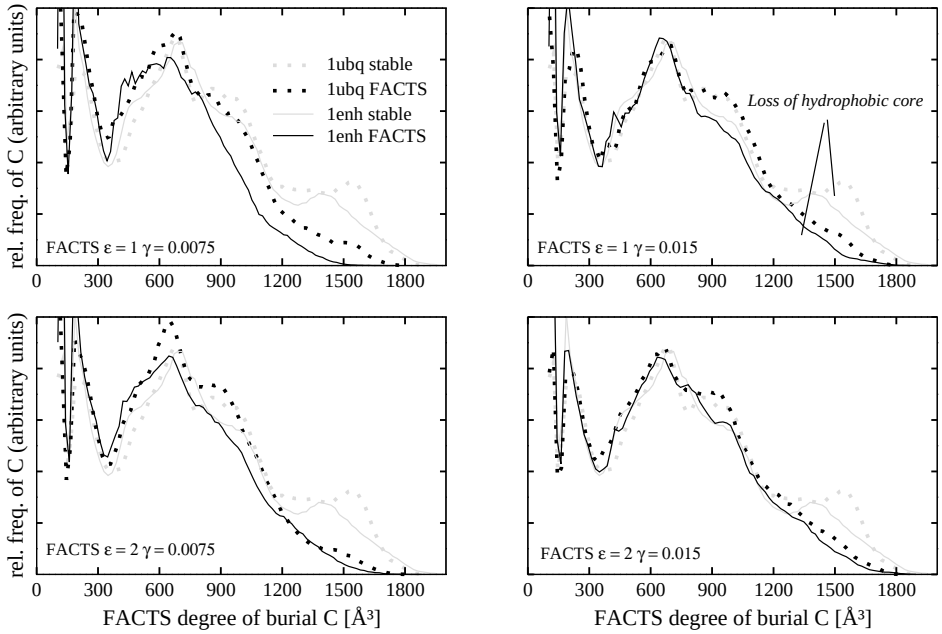


Figure 6.10: Loss of the hydrophobic core of 1ubq and 1enh in relation with the same FACTS simulations as in Fig. 9.91, seen by means the C spectra (the solid lines correspond to stable trajectories of 100 ns with harmonic constraints, while the dotted ones correspond to FACTS). The peaks around $C = 1500$ for 1ubq and $C = 1300$ for 1enh is lost along the FACTS simulations. This is the region in which the SISI nonpolar correction takes place (see Fig. 6.5).

6.4 Conclusions

On the one hand FACTS shows good behaviour with well solvated structures such as wkqa and act2, at least in comparison with other solvation models, but fails to achieve stability for more globular proteins like 1ubq, unless the parametrisation is changed. On the other hand FACTS SISI, which consists of a correction to nonpolar solvation energy based on the Tolman correction to SASA theory and on the original FACTS degree of burial of each atom, gets rid of this issue: peptide solvation features and 100-ns stability of the testcase proteins are provided within the same parameter setup. Limits to these tests are certainly due to the use of CHARMM param19. Future developments of FACTS SISI towards a full-atom force field are thus encouraged by this work.

This result confirms recent ideas [2, 3, 4, 5, 6, 7, 8, 9] about the treatment of

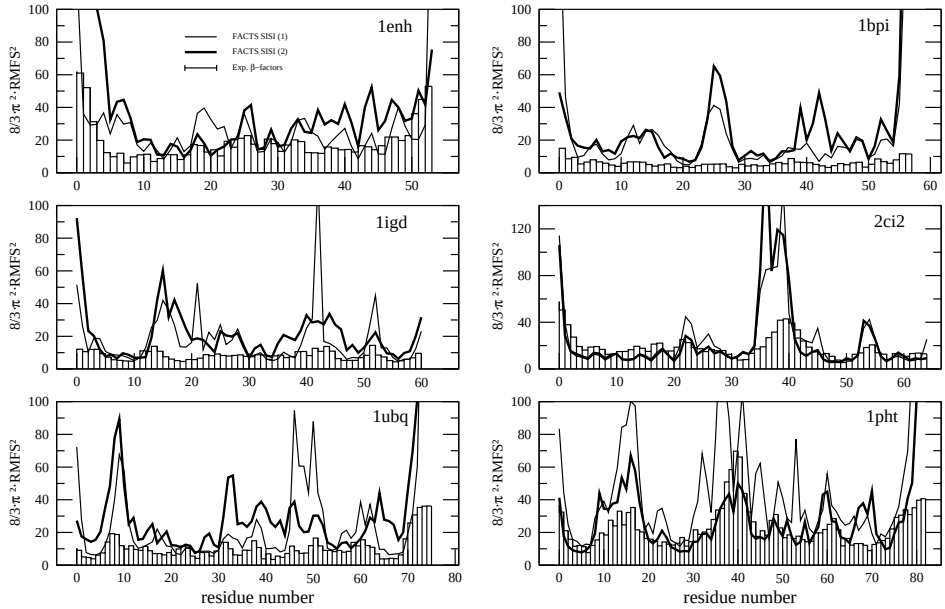


Figure 6.11: To ensure that the MD stability with FACTS SISI at 300 K illustrated in Tab. 6.5 is not due to overstabilization, a comparison between RMSD fluctuations and experimental protection factors (for 1enh, 1bpi 1igd, 2ci2, 1ubq and 1pht) has been performed (red and black lines are related to two different runs). MD peaks position and magnitude are in agreement with experimental data.

nonpolar interactions claiming that SASA approximation is not precise enough for an accurate treatment of biological molecules in their aqueous environment.

Chapter 7

Conclusions and future work

FACTS is fast and accurate, but the best parametrisation has to be selected according to the structure in study. FACTS SISI solves this problem, introducing only 3 additional parameters. The number of parameters could be further reduced by limiting the number of atom types. In future, it would be convenient to setup FACTS SISI in a full-atom model.

You can never solve a problem on the level on which it was created.
A. Einstein

Conclusion about the FACTS model. The popular implicit solvation models that can be efficiently employed for computer simulations are based on a distance–dependent screening function, rather than on a constant dielectric in the denominator of the Coulomb formula. However, the methods based on the Generalised-Born equation (including FACTS) approximate the screening effects by taking into account not only the charge–charge distance but also the degree of solvent exposure of individual charges. It is important to note that a direct comparison of the reliability of Generalised-Born approaches (including FACTS) to simple models based on distance–dependent screening function is not possible because of the different level of physical information and different number of parameters. There are also important differences in the evaluation of atomic (or self) solvation energy values between FACTS and other efficient implicit models. The Gaussian solvent–exclusion model of Lazaridis and Karplus EEF1 [71], the screened Coulomb potential (SCP) model of Hassan [48] and the recent ABSINTH con-

tinuum solvation model developed by Vitalis and Pappu [72] do not take into account *the spatial symmetry of the displaced solvent* whereas a symmetry term is explicitely used in FACTS. Furthermore, EEF1 assumes that the solvation free energy of a protein is a sum of group contributions and is parameterised with experimental data of small model compounds, whereas atomic solvation energies evaluate by finite different Poisson method are used in the parametrisation of FACTS. Hence, the EEF1 solvation energy cannot be decomposed into polar and nonpolar contributions.

Compared to most Generalised-Born models, a common advantage of FACTS, EEF1 and SCP is that they do not require the definition of a boundary between solute and solvent. On the other hand, the values of the dielectric constant of solute and solvent have to be specified in FACTS (and GB models) but not in EEF1 and SCP. Atomic solvation energies strongly depend on the dielectric constant of the solute ϵ. Yet, the possibility of defining a solute dielectric constant increases the range of applicability of FACTS because $\epsilon = 1$ is more appropriate for molecular dynamics simulations, while for structure prediction or docking values of $\epsilon = 2$ or $\epsilon = 4$ better approximate the effects of fluctuating dipoles in single-point energy calculations.

Conclusion about FACTS SISI model. The molecular dynamics simulations performed on the same testcase used to test FACTS showed that the introduction of the simple sigmoidal (SISI) correction into the treatment of nonpolar interactions enhances the stability of the proteins and, at the same time, preserves the good results with more unstructured peptides within the same parametrisation (in particular, $\epsilon = 2$, $\gamma = 7.5$ [cal $\cdot$ mol$^{-1}\cdot$ Å^{-2}]). The comparison between the results of chapter 4 and chapter 6 allows us to appreciate the relevance of the SISI correction to FACTS. As pointed out in chapter 3, the correction can be interpreted as a combination between effects intrinsic to the FACTS model (essentially, the tendency of the original version of FACTS to overestimate solvation energy for buried atoms) and extrinsic (nonpolar treatment of solvation energy via Tolman's theory).

Further developments of FACTS SISI should implement the SISI nonpolar correction into an all-atom forcefield such as CHARMM22. Conversely, attempts to correct the original FACTS *geometric definition* of A and B or even improvements in the fitting

procedure with fdP seem not to be useful, since highly-enhanced geometry did not give significantly better results (from the atomic solvation energy point of view). On the other hand, it is crucial *to limit the number of parameters* of such a corrected model. Let us summarise which parameters are involved in the FACTS SISI model setup. FACTS fit parameters, 9 for each atom type from FACTS setup – R^{sphere} and 4 sigmoidal coefficient in the definition of electrostatic setup plus 4 from the definition of the SASA (nonpolar) setup – result in 63 parameters for CHARMM19 (7 atom types) and 171 for CHARMM22 (19 atom types). FACTS SISI includes 3 more parameters, necessary to setting up the curvature correction. Finally, we have always to take into account the two FACTS free parameters (internal dielectrics ϵ for electrostatics and surface tension γ from the SASA approximation of nonpolar interactions). To sum up, **63+5** parameters for CHARMM19 and **171+5** parameters for CHARMM22. This large number of parameters causes difficulties in the development of the model, and exposes the model to overfitting. Indeed, the SISI correction was purposely designed in the spirit of limiting the growth in the number of parameters.

In chapter 6 it has been shown that the trend of the discrepancy between Tolman and SASA theory has a strong dependence on the atom type, if studied as a function of the S_i (the contribution to the solvent accessible surface area of atom i) or the σ_i (the curvature of the S_i surface). But if studied as a function of the FACTS degree of burial C_i, these differences in the trend disappear. The trend is actually a sigmoidal function for each atom type. This allowed us to average these parameters and set a unique function (the SISI function) valid for all the atom types, reducing the (new) parameters from 21 to 3. Regarding the fitting procedure, a possible way to reduce the number of parameters could consist in reducing the atom type number to 2 atom types (hydrogen and not-hydrogen atoms). As one can argue from chapter 3 and from the error analysis in chapter 5, the main differences among the double sigmoidal fitting functions (for the electrostatics and the SASA terms) are actually due to the atom type H and the others (atom type S can be simply neglected). Hence, the model will be significantly lighter ($9 \times 2 = 18$ instead of 63 fitting parameters for CHARMM19 and $9 \times 2 = 18$ fitting parameters instead of 171 for CHARMM22).

Chapter 8

Acknowledgements

Friendship does not solve problems. It helps not to afford them alone.
P. Manco

First and foremost I am grateful to those who helped me complete this work. To **A. Caflisch**, for putting his trust in me and allowing me to work in the Universität Zürich and to all the people in the group: **G. Interlandi** (he was my first friend in Zürich and gave me important Überlebentipps for living in this town, like *La Scala*), **A. Cavalli** (who was patient, even when I asked him several times about the meaning of `rm -rf * /`), **B. Paoli** (for her great heart, for her strong pragmatism and for the capacity we had to solve problems together, even if bickering), **R. Pellarin** (for his great knowledge of this complex discipline, his help and his suggestions), **R. Scalco** (for keeping awake the importance of Physics in my life, for he shared with me his last, unique *Nardini*), **P. Schütz** (for he is still the strongest, the Captain, and for his indestructible helpfulness), **F. Marchand** (because he is the only other lucky guy who knows the subtle joy of parametrising FACTS with CHARMM, param22), **R. Curcio** (for his encyclopedic knowledge of CHARMM), **F. Dey** (for his typical berner understatement, for his skills in "cluster queueing management", whatever this means), **E. Guarnera** (for our discussions about the best way to solve the biggest problems in the world, while having both particularly great difficulties in afford the little problems of our own lives); **P. Alfarano** (for he is the best programmer I have ever met, for his irresistible, unmistakable sense of humour, and also for being so impatient but willing, snob and generous, kindly unpleasant, cynic and dreamer),

I would like also to say *merci* to **C. Gujan** for her support with burocracy and languages, and for having shared with her the everyday life in the office. *Je te souhaite un bon weekend encore, Christiane.* Furthermore I am very grateful to the mathematicians: **E. Spadaro** (for his bright, deep soul, for his presence in the distance, his sympathy disguised with irony, and for the gift of sharing his friendship), **F. Maganiello** (for he had always a good word for me and he was always ready to help me - and everybody else). Thanks also to **F. de Luca** (for the past days, far away) and to **M. Lavagnini** (for her patience, her joy, her sweet heart and her piercing, perceptive mind). Und **A. Bičanski**, vielen Dank für die Freundschaft, für die Nähe, für Lausanne und für die Korrektur dieser Arbeit.

My greatest thanks go eventually to *mamma e papà* and to my `frater` Marco and to Alessia (sorry if I was not there in these years) and also to Patrizio e Valentina Caricato, to Pier, to Nicola, to Remo, to Gerry, to Arianna and to Eva, my so dear, far-off friends, distant only in a metrical sense, but nicely, intimately and topologically merged to my life.

Chapter 9

Appendix

9.1 Appendix 1

9.1.1 Supplementary Material (chapter 4)

From Results and Discussion

Shifts distributions recovered via SHIFTX in this work are usually unimodal and they are well fitted by a single gaussian peak: the standard deviation of these distributions is assumed as the error in the CS prediction for unimodal distribution (σCS). Once σCS is known for each fit, the comparison with experimental data is performed by means of the χ^2 variable

$$\chi^2 = \sum_i^N (CS_i^{exp} - CS_i^{sim})^2 / \sigma CS_i^2,$$

where i counts the available shifts, since the number of studied shifts N (corresponding to the number of degree of freedom (DF) of the variable) is always greater than 15 (the fit were performed with Gnuplot 4.2). Experimental CS will be removed from the analysis if their differences between the calculated ones will exceed 3σCS in each parameterisation. Since the ranges of variation of HCβ, HCβ_2 and HCβ_3 shifts are similar, they will be grouped in the same statistical sample in order to be compared with experimental values.

PMF between charged sidechains with FACTS

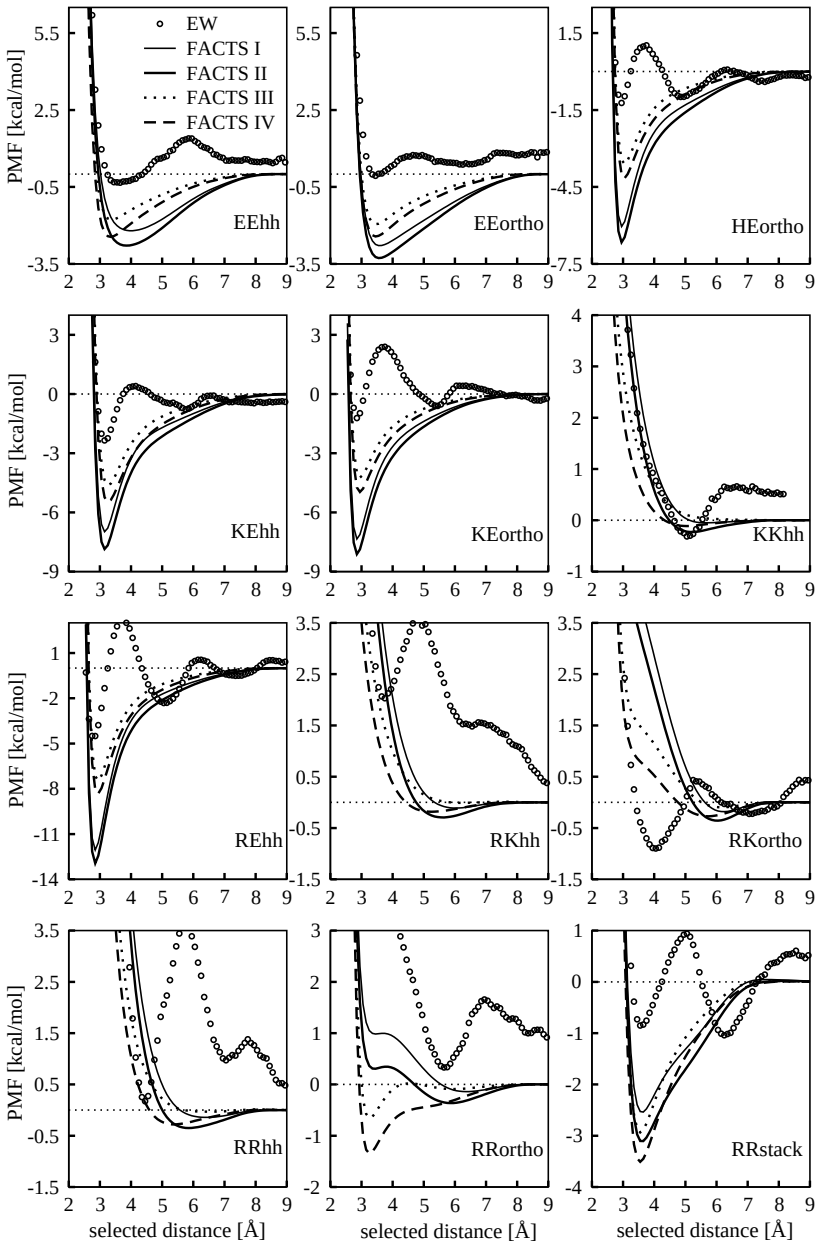


Figure 9.1: Explicit-water PMF for ionisable sidechains as published in Ref. [30] compared with different FACTS setups. For a more quantitative analysis, a comparison between the first minimun in the PMF profile has been done (Tab. 9.1- 9.2). FACTS IV ($\epsilon = 2$ and $\gamma = 7.5$) results closest to explicit water results.

app.	EW	GBMV	GBSW	EEF1	FACTS I	FACTS II	FACTS III	FACTS IV
EEhh	**-0.33**	0.48	0.31	0.04	-2.21	-2.79	-1.78	-2.45
EEor	**-0.06**	0.47	-0.05	-2.38	-2.79	-3.27	-1.95	-2.43
HEor	**-1.22**	-2.49	-1.82	-	-6.05	-6.66	-3.61	-4.16
Kehh	**-2.35**	-1.62	-2.04	-2.26	-6.97	-7.86	-4.62	-5.48
Keor	**-1.22**	-2.48	-1.69	-6.09	-7.35	-8.12	-4.24	-4.96
KKhh	**-0.32**	0.60	0.38	0.05	-0.05	-0.23	-0.00	-0.11
REhh	**-4.50**	-3.90	-2.54	-8.71	-12.06	-12.94	-7.46	-8.32
RKhh	**0.32**	0.39	-0.02	0.03	-0.11	-0.29	-0.01	-0.18
RKor	**-0.90**	0.51	-0.68	-1.16	-0.18	-0.36	-0.10	-0.27
RRhh	**0.17**	0.53	-0.46	0.07	-0.14	-0.35	-0.03	-0.28
RRor	**0.33**	0.35	-1.77	-8.06	-0.14	-0.36	-0.65	-1.35
RRst	**-1.04**	0.06	-3.42	0.15	-2.54	-3.10	-2.93	-3.50

Table 9.1: Absolute value (kcal/mol) of the first minimum of the PMF related to ionisable sidechains approach (app.) shown in Fig 9.1 with different solvation models and FACTS with four different parameter sets. Capital letters refer to amino acid code, while hh=head to head approach, or=orthogonal approach; st=stacked approach, as indicated in Ref. [30].

app.	I dev.	II dev.	III dev.	IV dev.
EEhh	1.88	2.46	1.45	2.12
EEor	2.73	3.21	1.89	2.37
HEor	4.83	5.44	2.39	2.94
Kehh	4.62	5.51	2.27	3.13
Keor	6.13	6.90	3.02	3.74
KKhh	0.27	0.09	0.32	0.21
REhh	7.56	8.44	2.96	3.82
RKhh	0.43	0.61	0.33	0.50
RKor	0.72	0.54	0.80	0.63
RRhh	0.31	0.52	0.20	0.45
RRor	0.47	0.69	0.98	1.68
RRst	1.50	2.06	1.89	2.46
tot.	*31.45*	*36.47*	*18.5*	*24.05*

Table 9.2: Absolute values (kcal/mol) of the difference between EW minima in Tab. 9.1 and the FACTS results. Simulations with $\epsilon = 2$ and $\gamma = 7.5$ (FACTS III) are the best parameter set approximating the first minimum of the Lazaridis PMF.

Tyrosine hydroxylase

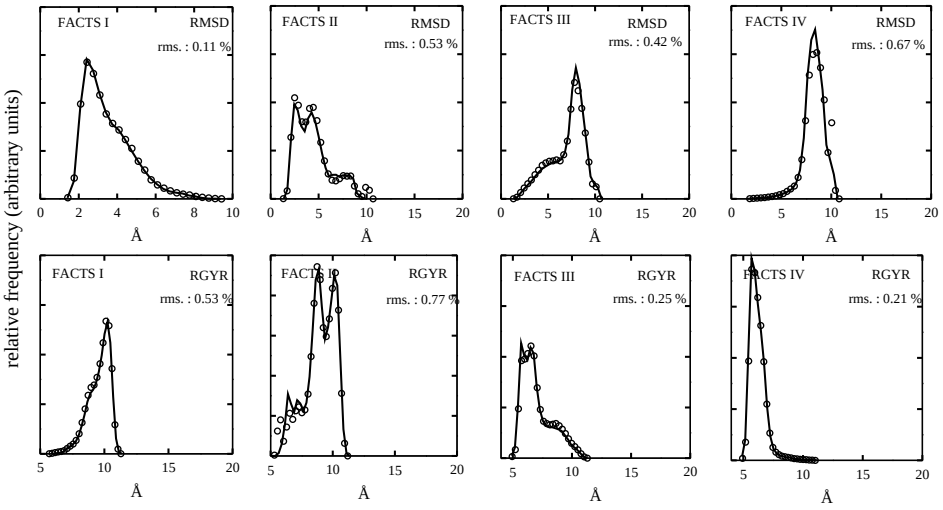


Figure 9.2: Convergence test based on halving the 4 μs long wkqa simulations with FACTS (at 300 K) in two sections. The reference structure for the RMSD timeseries is an extended structure. Circles refer to the first section of the trajectory; solid lines refer to the second half (the % deviation between the RMSD (and RGYR) distributions in the first and in the second half is shown).

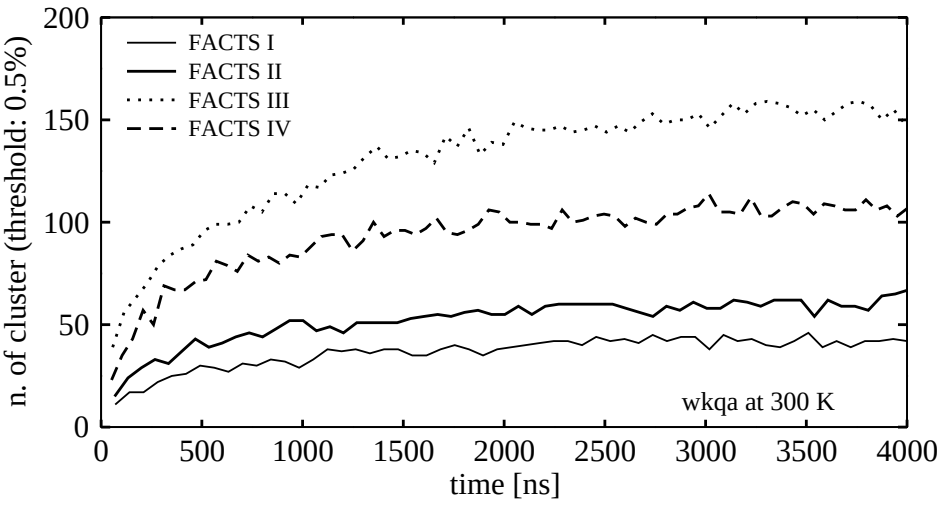


Figure 9.3: Convergence test for the same simulations as in Fig. 9.2 based on the number of significantly populated clusters.

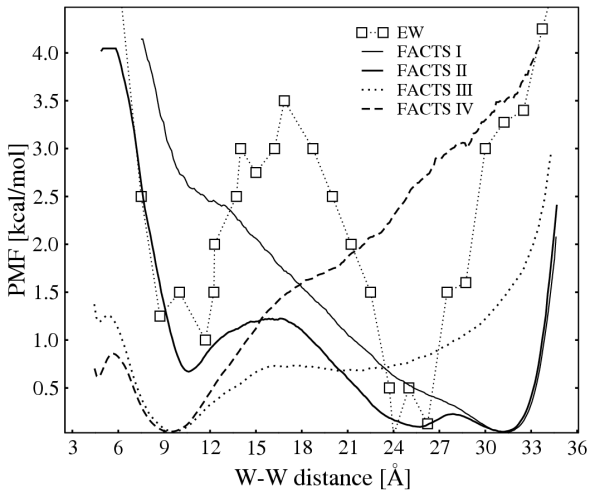


Figure 9.4: End-to-end-distance PMF force of wkqa at 300 K with different FACTS setups in comparison with Stultz's explicit water data (square-dotted line) from Ref. [32]. Distance is taken between Cα atoms of each tryptophan.

Melittin

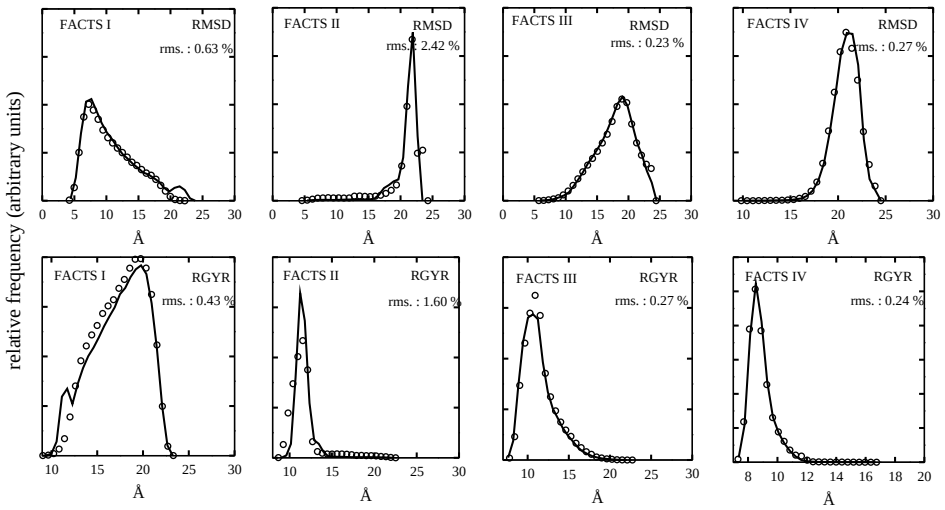


Figure 9.5: Convergence test based on halving the 6 μs long meli simulations with FACTS (at 303 K) in two sections. The reference structure for the RMSD timeseries is 1mlt. Circles refer to the first section of the trajectory; solid lines refer to the second half (the % deviation between the RMSD (and RGYR) distributions in the first and in the second half is shown).

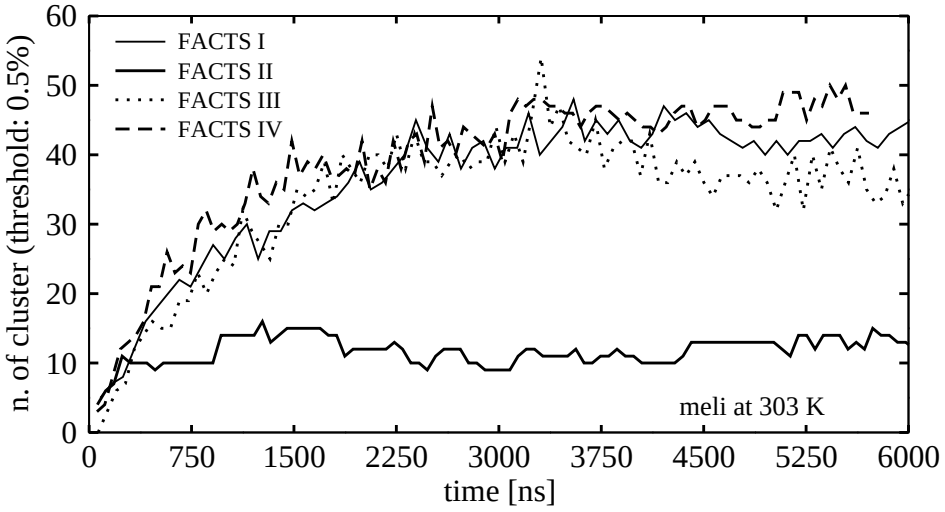
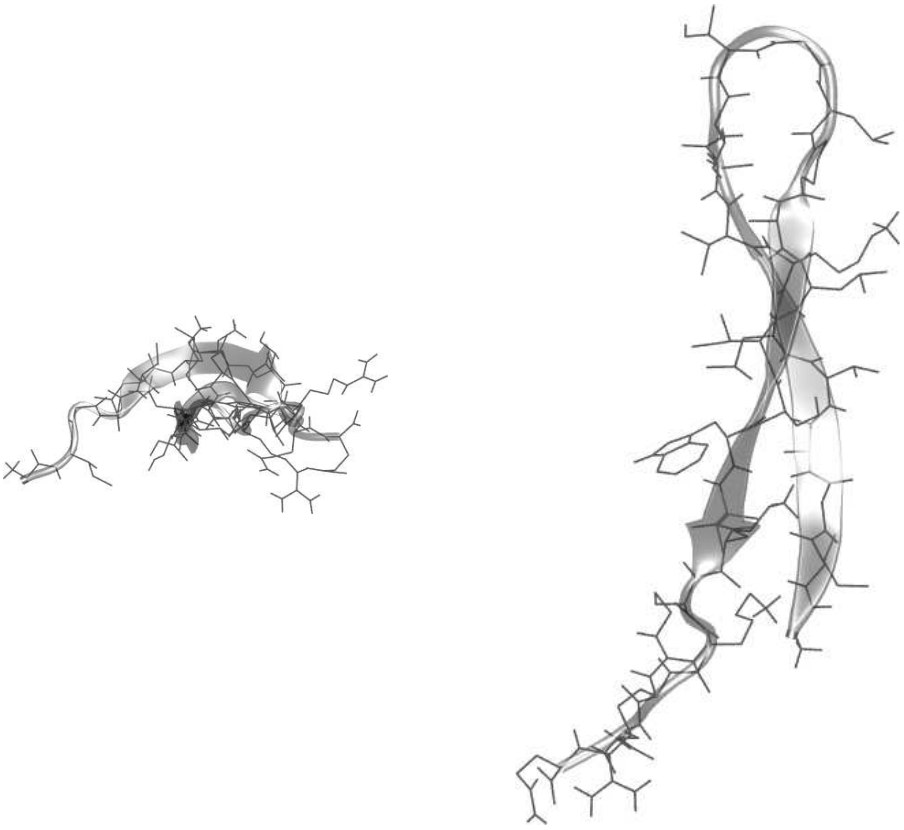
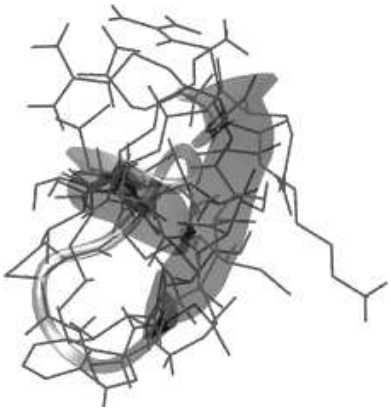


Figure 9.6: Convergence test for the same simulations as in Fig. 9.5 based on the number of significantly populated clusters.

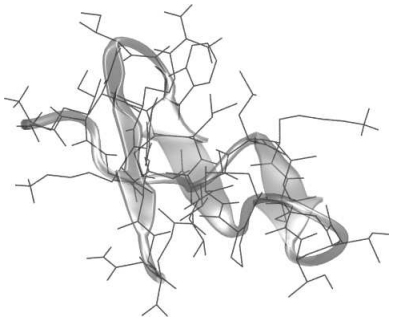


FACTS I (pop. 0.48 %)

FACTS II (pop. 16.32 %)



FACTS III (pop. 0.78 %)



FACTS IV (pop. 3.32 %)

Figure 9.7: Central conformations of peptide meli (at 300 K). The RMSD-clustering was performed with Wordom with a cutoff of 2.5 . Simulations are 6 μs long.

β-hairpin of protein G

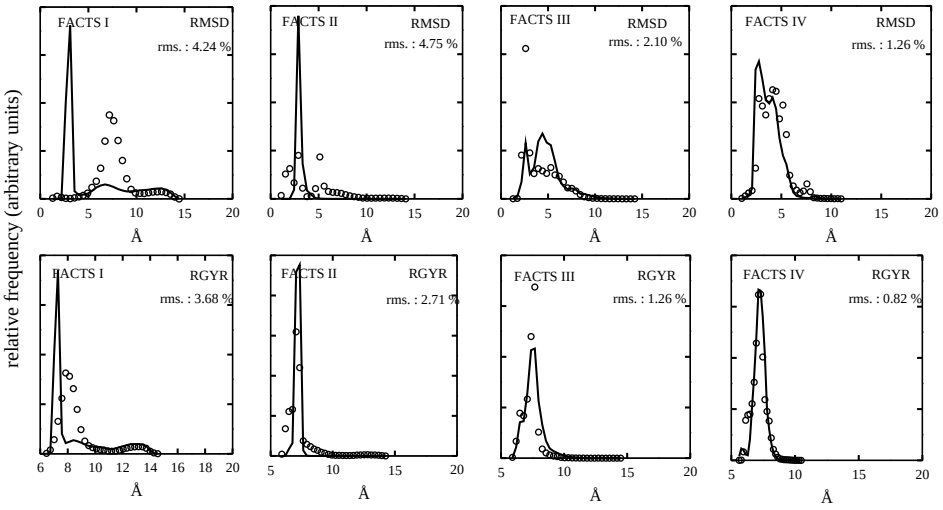


Figure 9.8: Convergence test based on halving the 4 μs long pgbh simulations with FACTS (at 280 K) in two sections. The reference structure for the RMSD timeseries is 1pgb. Circles refer to the first section of the trajectory; solid lines refer to the second half (the % deviation between the RMSD (and RGYR) distributions in the first and in the second half is shown). Even after 4 μs, the peptide seems not to be at equilibrium.

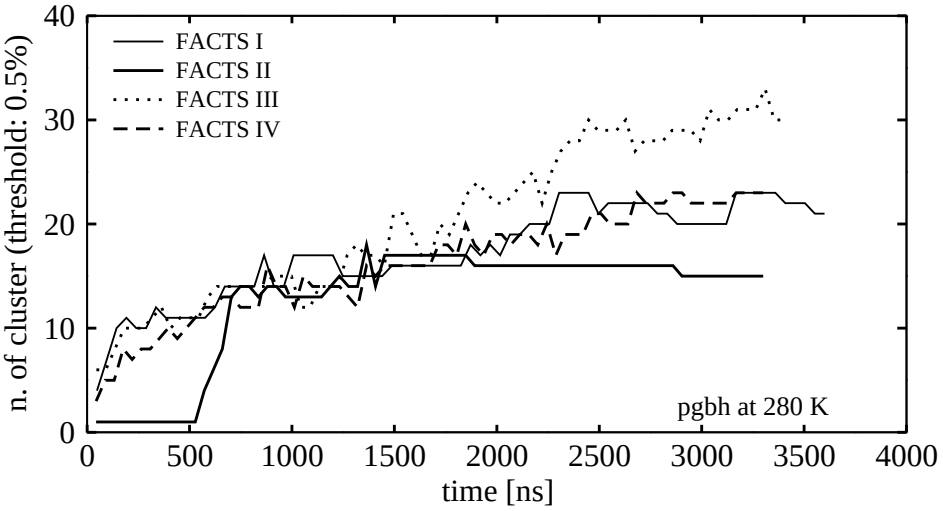
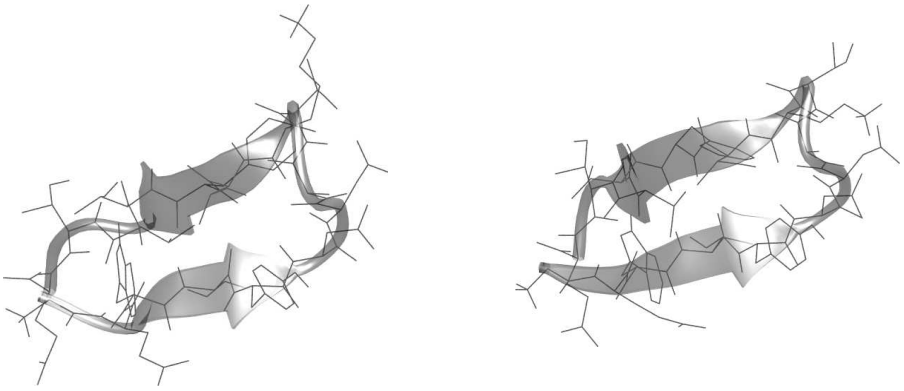
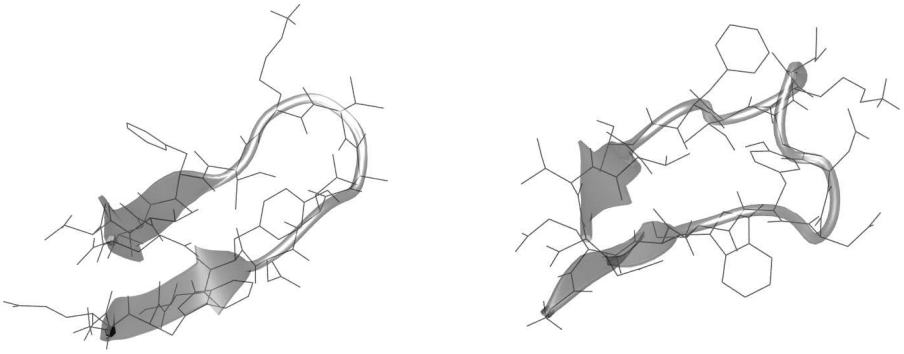


Figure 9.9: Convergence test for the same simulations as in Fig. 9.8 based on the number of significantly populated clusters. This plot confirm the hypothesis of non fully equilibrated simulation.



FACTS I (pop. 28.48 %) FACTS II (pop. 55.88 %)



FACTS III (pop. 16.02 %) FACTS IV (pop. 8.64 %)

Figure 9.10: Central conformations of peptide pgbh (at 280 K). The RMSD-clustering was performed with Wordom with a cutoff of 2.5 . Simulations are 4 μs long.

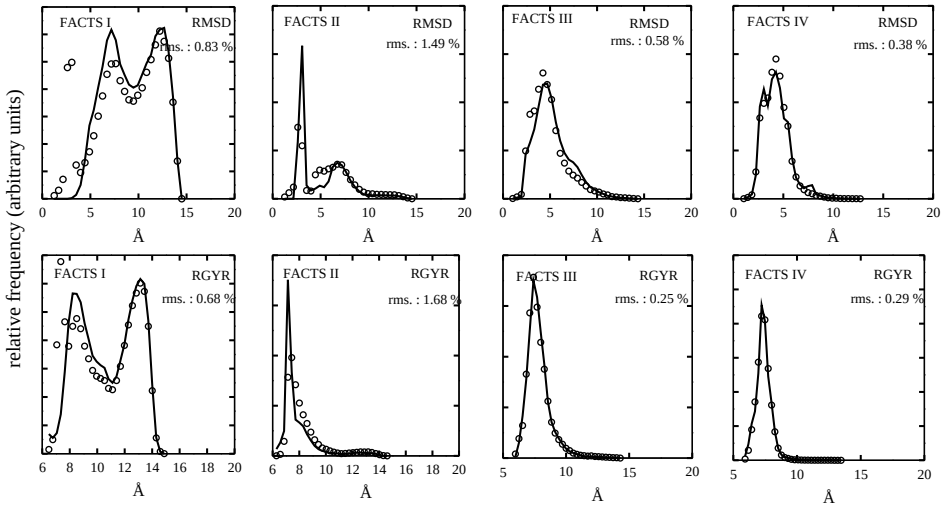


Figure 9.11: Convergence test based on halving the 4 μs long pgbh simulations with FACTS (at 300 K) in two sections. The reference structure for the RMSD timeseries is 1pgb. Circles refer to the first section of the trajectory; solid lines refer to the second half (the % deviation between the RMSD (and RGYR) distributions in the first and in the second half is shown).

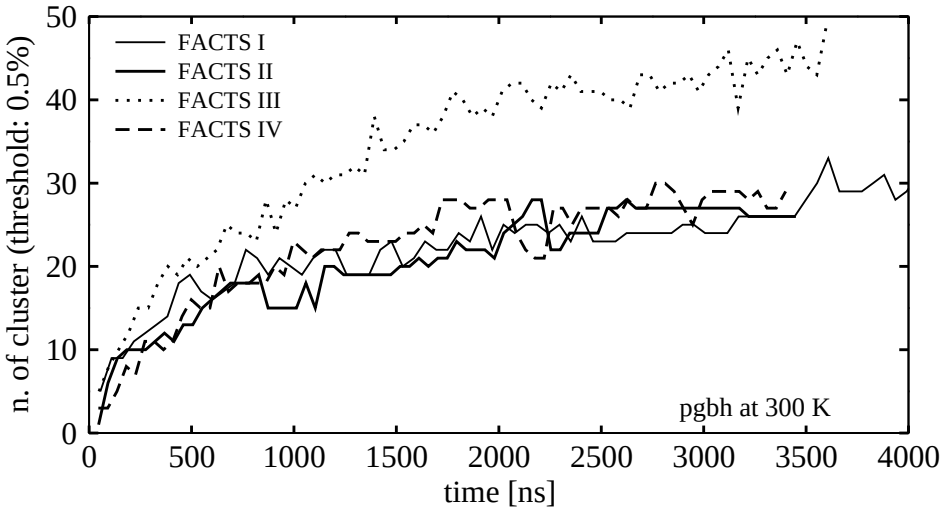
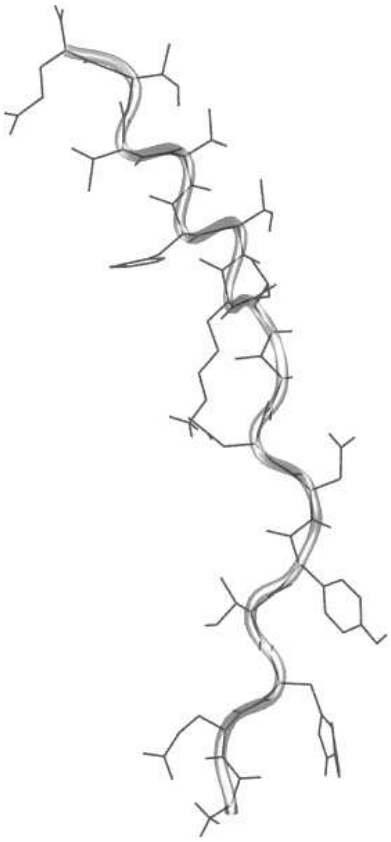
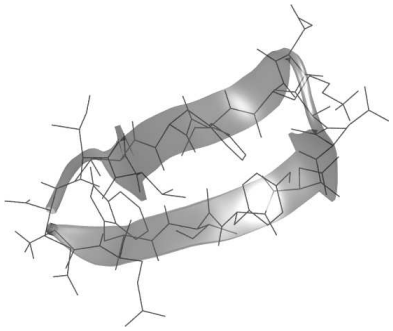


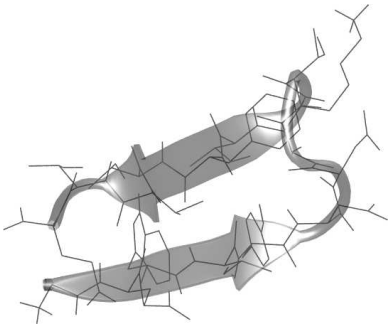
Figure 9.12: Convergence test for the same simulations as in Fig. 9.11 based on the number of significantly populated clusters.



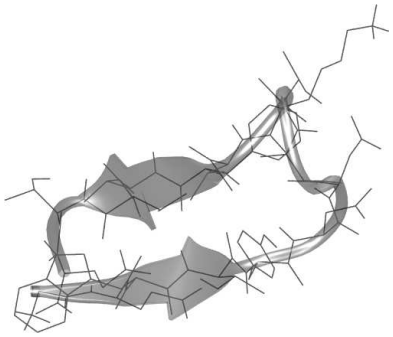
FACTS I (pop. 6.68 %)



FACTS II (pop. 27.91 %)



FACTS III (pop. 6.50 %)



FACTS IV (pop. 5.68 %)

Figure 9.13: Central conformations of peptide pgbh (at 300 K). The RMSD-clustering was performed with Wordom with a cutoff of 2.5 . Simulations are 4 μs long.

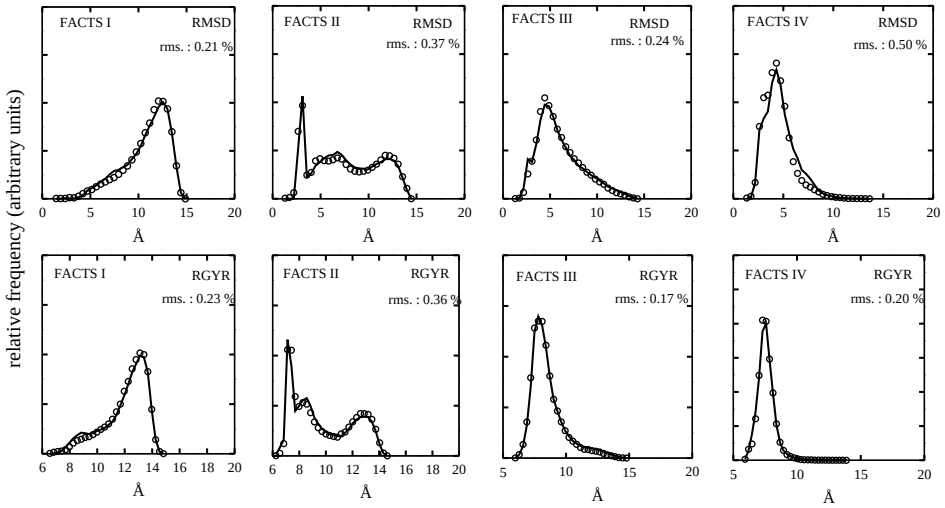


Figure 9.14: Convergence test based on halving the 4 μs long pgbh simulations with FACTS (at 320 K) in two sections. The reference structure for the RMSD timeseries is 1pgb. Circles refer to the first section of the trajectory; solid lines refer to the second half (the % deviation between the RMSD (and RGYR) distributions in the first and in the second half is shown).

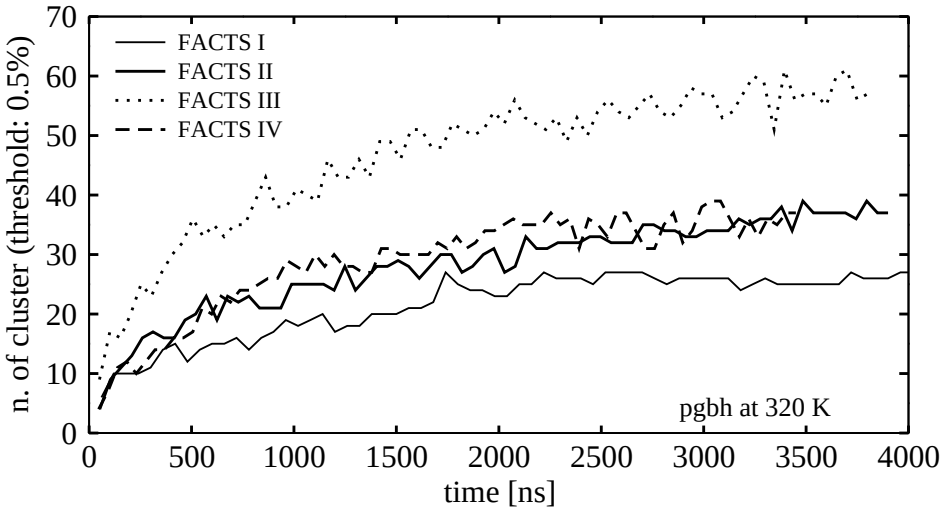
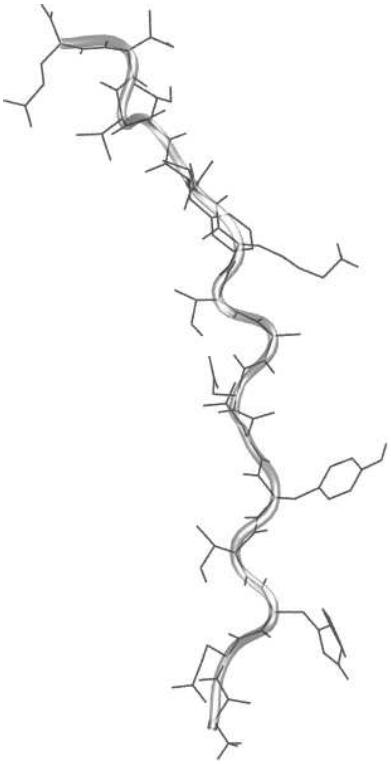
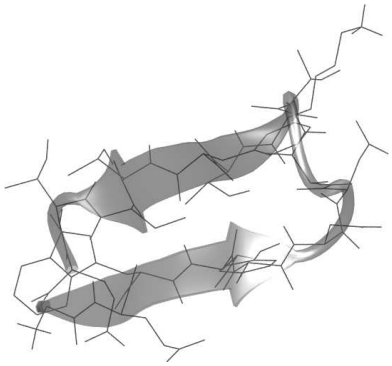


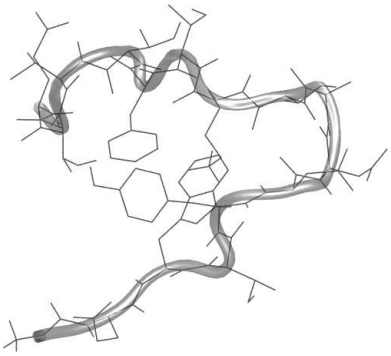
Figure 9.15: Convergence test for the same simulations as in Fig. 9.14 based on the number of significantly populated clusters.



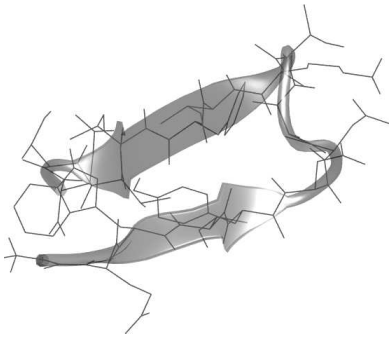
FACTS I (pop. 10.78 %)



FACTS II (pop. 15.28 %)



FACTS III (pop. 4.09 %)



FACTS IV (pop. 6.69 %)

Figure 9.16: Central conformations of peptide pgbh (at 320 K). The RMSD-clustering was performed with Wordom with a cutoff of 2.5 . Simulations are 4 μs long.

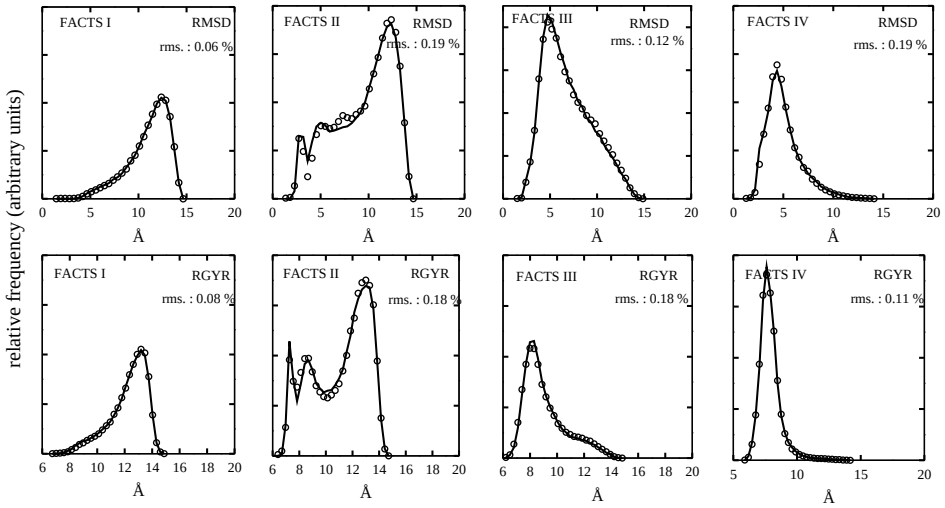


Figure 9.17: Convergence test based on halving the 4 μs long pgbh simulations with FACTS (at 340 K) in two sections. The reference structure for the RMSD timeseries is 1pgb. Circles refer to the first section of the trajectory; solid lines refer to the second half (the % deviation between the RMSD (and RGYR) distributions in the first and in the second half is shown).

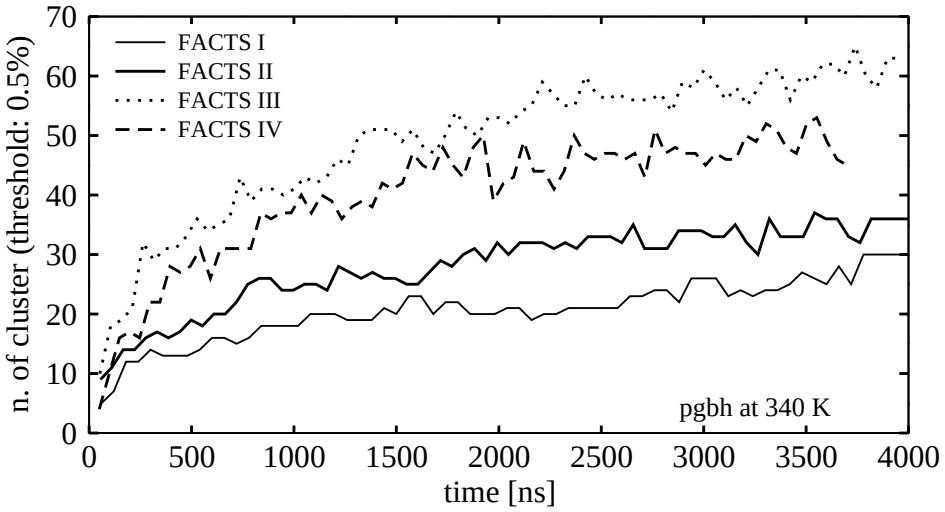
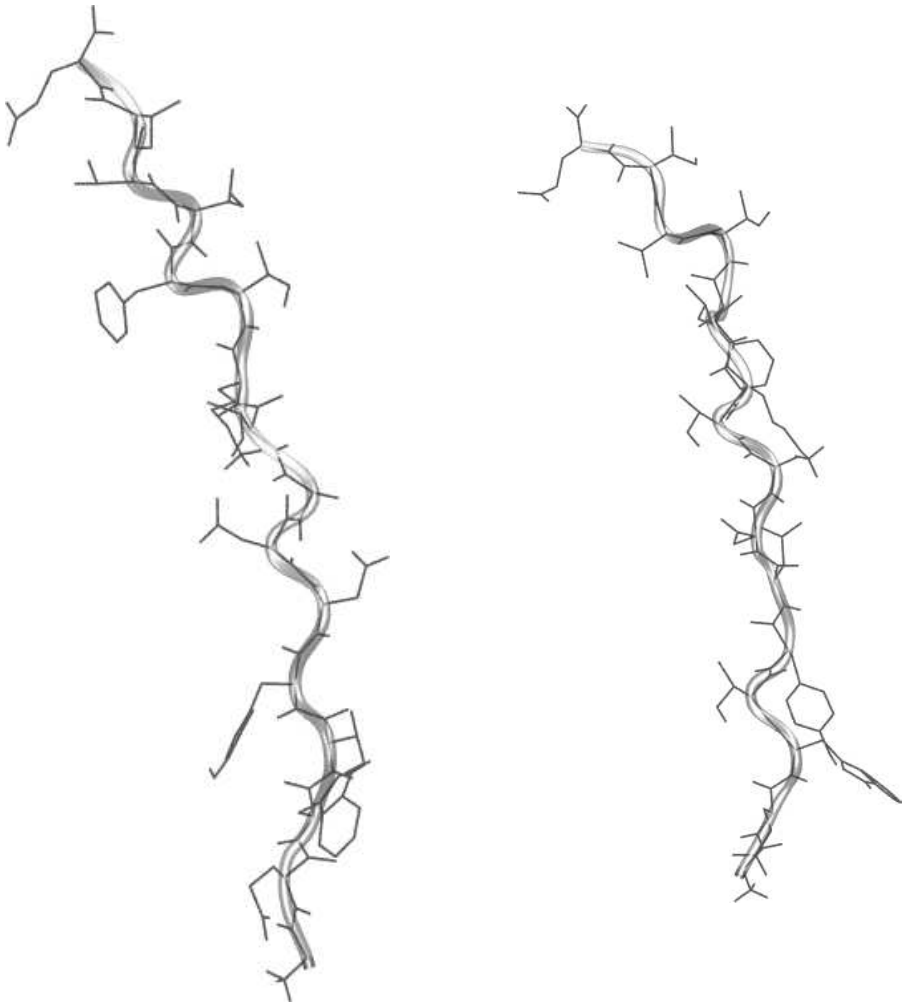
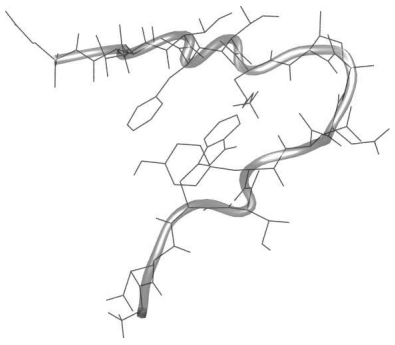


Figure 9.18: Convergence test for the same simulations as in Fig. 9.17 based on the number of significantly populated clusters.

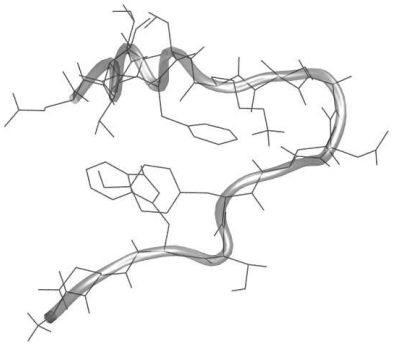


FACTS I (pop. 11.28 %)

FACTS II (pop. 7.50 %)



FACTS III (pop. 2.84 %)



FACTS IV (pop. 3.63 %)

Figure 9.19: Central conformations of peptide pgbh (at 340 K). The RMSD-clustering was performed with Wordom with a cutoff of 2.5 . Simulations are 4 μs long.

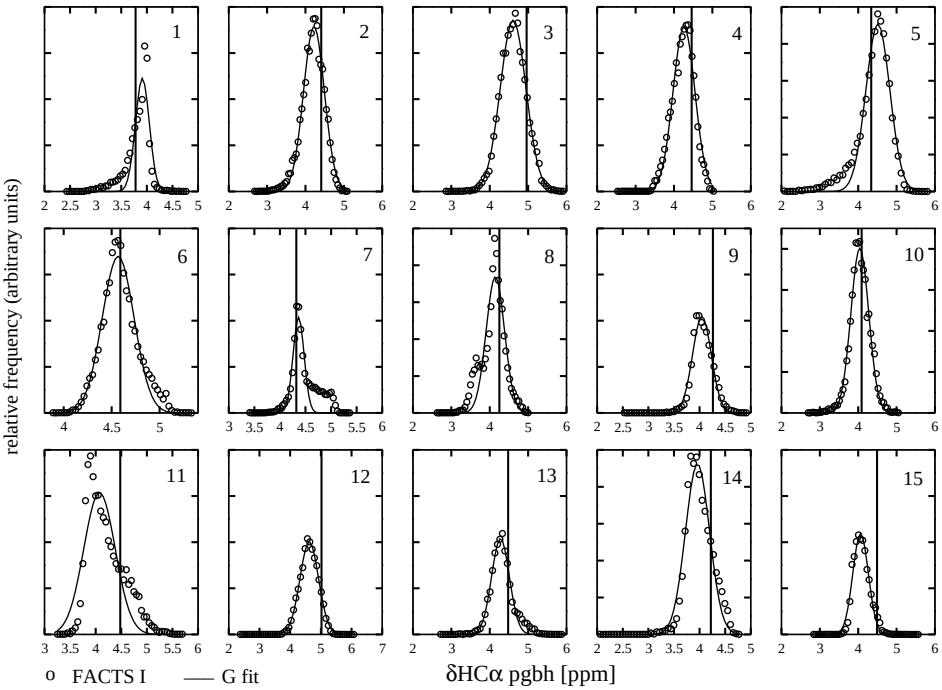


Figure 9.20: pgbh HCα CS calculated via SHIFTX program from the FACTS simulations (4 μs) at 280 K with FACTS I. Vertical lines represent the experimental shifts. See Tab. 9.3, Tab. 9.4 and Fig. 9.24 for quantitative analysis.

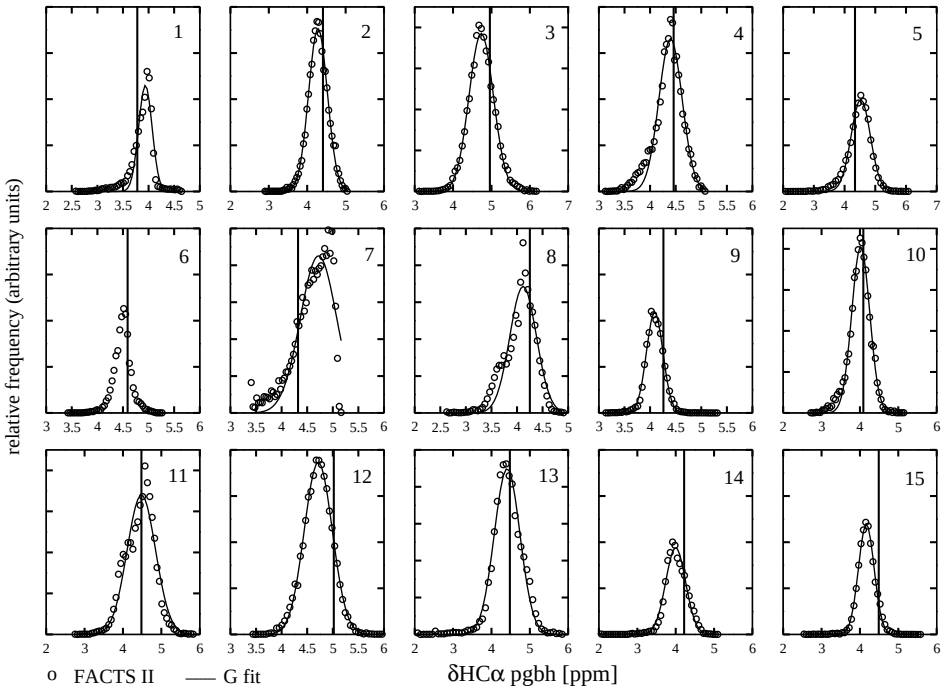


Figure 9.21: pgbh HCα CS calculated via SHIFTX program from the FACTS simulations (4 μs) at 280 K with FACTS II. Vertical lines represent the experimental shifts. See Tab. 9.3, Tab. 9.4 and Fig. 9.24 for quantitative analysis.

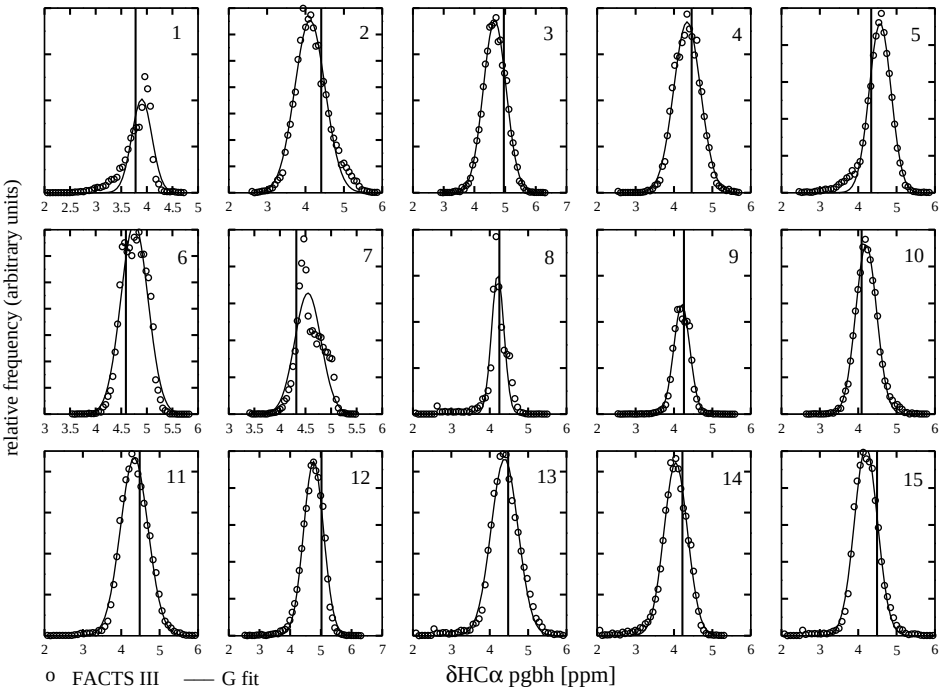


Figure 9.22: pgbh HCα CS calculated via SHIFTX program from the FACTS simulations (4 μs) at 280 K with FACTS III. Vertical lines represent the experimental shifts. See Tab. 9.3, Tab. 9.4 and Fig. 9.24 for quantitative analysis.

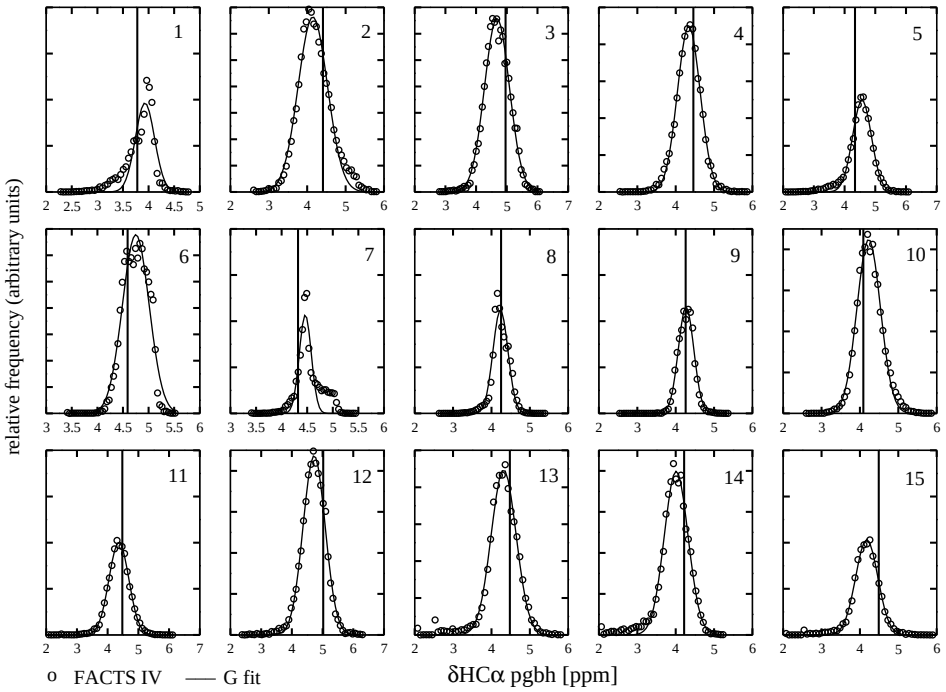


Figure 9.23: pgbh HCα CS calculated via SHIFTX program from the FACTS simulations (4 μs) at 280 K with FACTS IV. Vertical lines represent the experimental shifts. See Tab. 9.3, Tab. 9.4 and Fig. 9.24 for quantitative analysis.

res.	exp. HCα	FACTS sim. HCα	I σHCα	FACTS sim. HCα	II σHCα	FACTS sim. HCα	III σHCα	FACTS sim. HCα	IV σHCα
1	**3.78**	3.91	0.14	3.94	0.13	3.90	0.20	3.93	0.19
2	**4.41**	4.22	0.27	4.28	0.24	4.12	0.39	4.15	0.38
3	**4.96**	4.61	0.33	4.73	0.32	4.66	0.37	4.68	0.41
4	**4.46**	4.26	0.29	4.40	0.23	4.34	0.35	4.34	0.30
5	**4.34**	4.51	0.31	4.53	0.31	4.57	0.28	4.57	0.31
6	**4.59**	4.57	0.17	4.57	0.37	4.76	0.26	4.75	0.26
7	**4.32**	4.36	0.11	4.72	0.35	4.55	0.26	4.46	0.12
8	**4.25**	4.15	0.24	4.12	0.26	4.21	0.16	4.23	0.20
9	**4.26**	4.05	0.18	4.09	0.16	4.20	0.21	4.27	0.20
10	**4.09**	4.04	0.23	4.02	0.23	4.21	0.27	4.23	0.31
11	**4.48**	4.07	0.31	4.48	0.37	4.34	0.36	4.37	0.32
12	**5.02**	4.62	0.30	4.72	0.27	4.76	0.32	4.73	0.36
13	**4.48**	4.28	0.23	4.41	0.33	4.39	0.35	4.32	0.32
14	**4.22**	3.96	0.24	4.00	0.26	4.05	0.31	4.01	0.31
15	**4.49**	4.06	0.20	4.17	0.20	4.20	0.30	4.18	0.29
16	**4.13**	4.16	0.03	4.14	0.03	4.18	0.04	4.18	0.07

Table 9.3: Comparison between experimental (bold) and simulated δHCα CS of pgbh with FACTS (at 280 K).

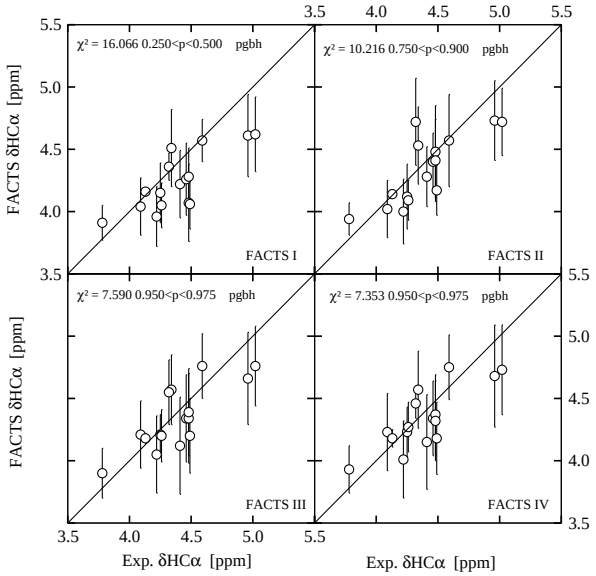


Figure 9.24: HCα CS of FACTS simulations with pgbh in comparison with experimental values (280 K).

par.	DF	χ^2	p
FACTS I	16	16.066	0.250¡p¡0.500
FACTS II	16	10.2157	0.750¡p¡0.900
FACTS III	16	7.58954	0.950¡p¡0.975
FACTS IV	16	7.35341	0.950¡p¡0.975

Table 9.4: Statistical analysis of pgbh δ HCα shifts (at 280 K) for experimental and calculated values.

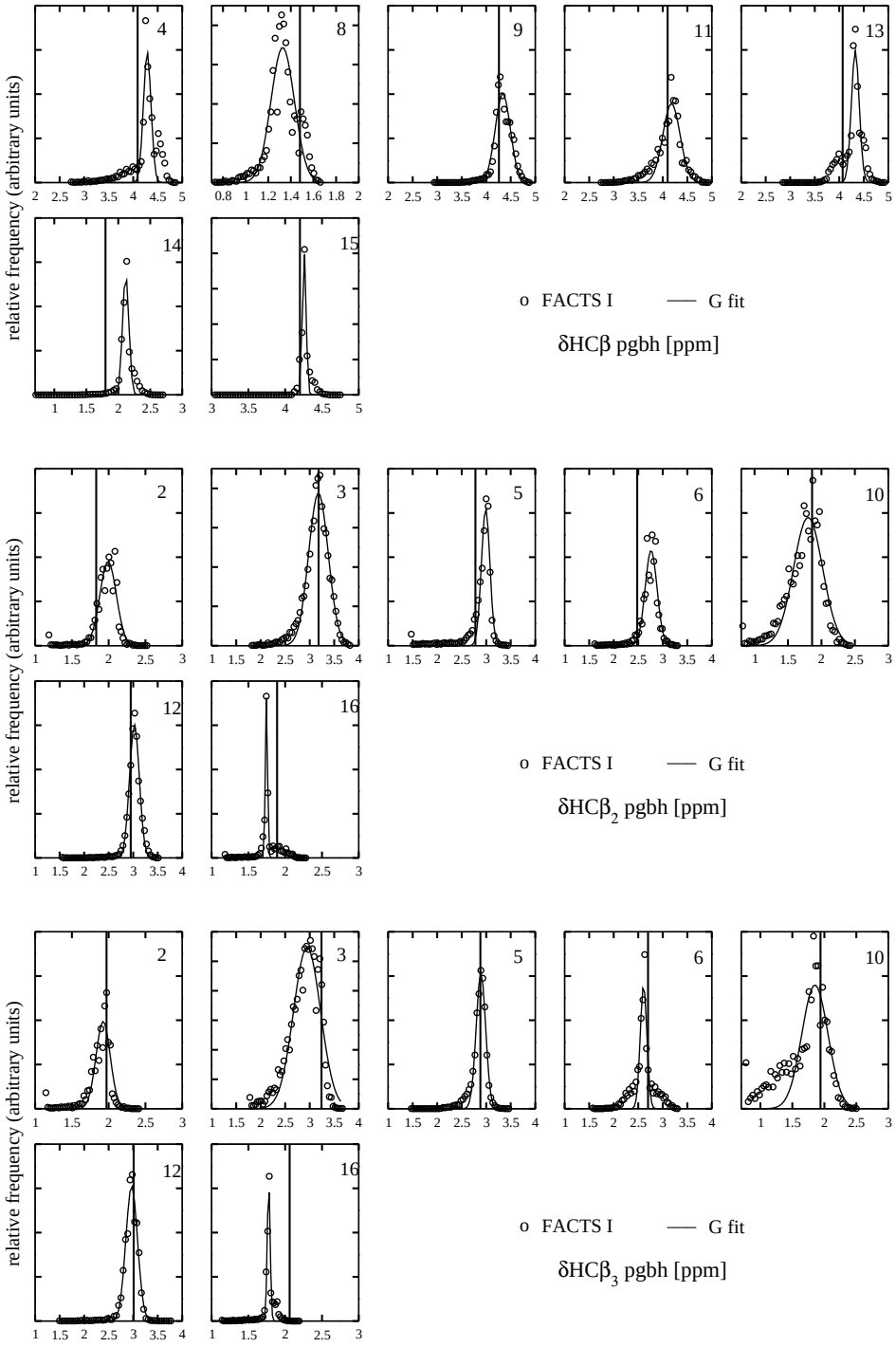


Figure 9.25: pgbh HCβ, HCβ_2 and HCβ_3 CS calculated via SHIFTX program from the FACTS simulations (4 μs) at 280 K with FACTS I. Vertical lines represent the experimental shifts. See Tab. 9.3, Tab. 9.4 and Fig. 9.24 for quantitative analysis.

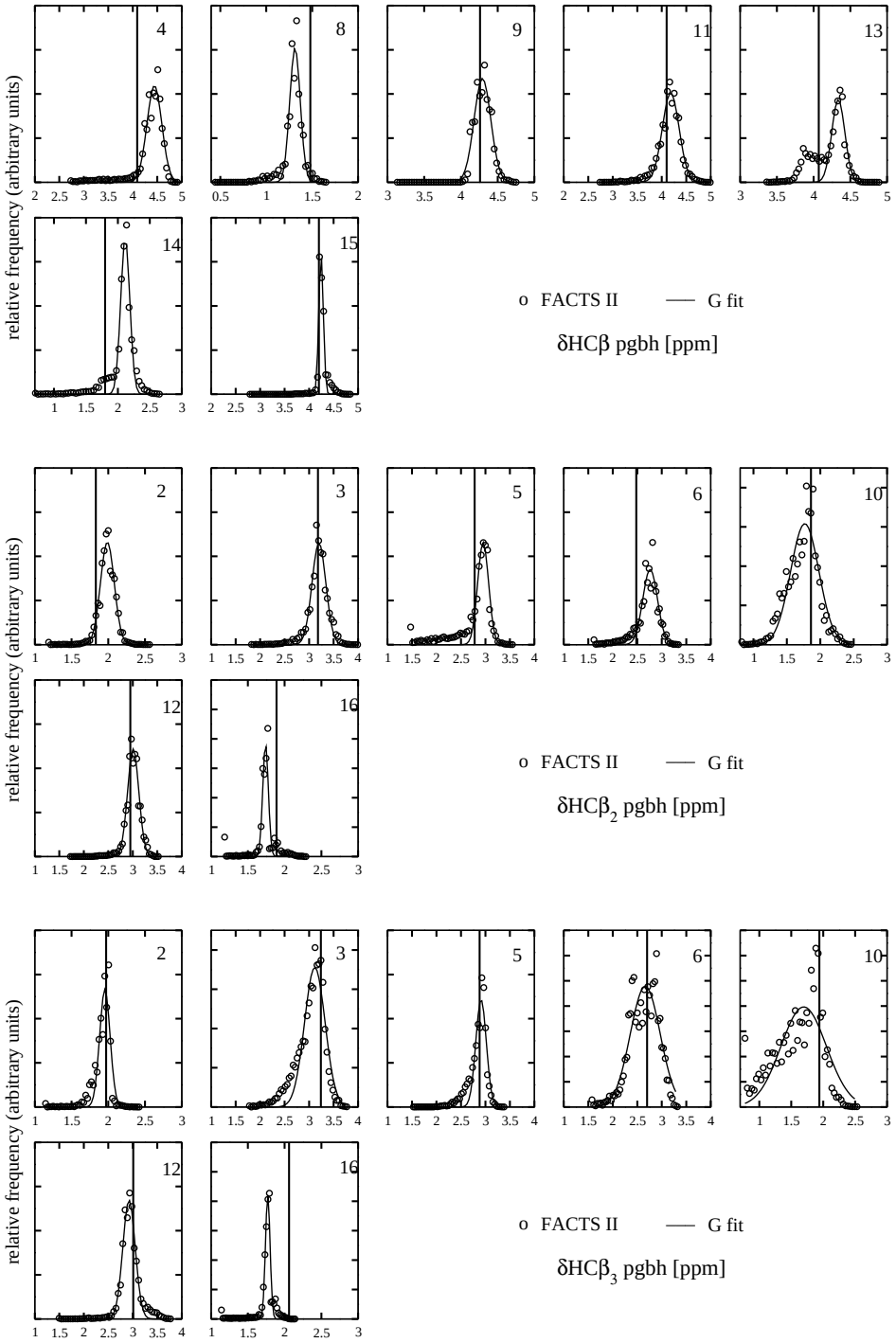


Figure 9.26: pgbh HCβ, HCβ_2 and HCβ_3 CS calculated via SHIFTX program from the FACTS simulations (4 μs) at 280 K with FACTS II. Vertical lines represent the experimental shifts. See Tab. 9.3, Tab. 9.4 and Fig. 9.24 for quantitative analysis.

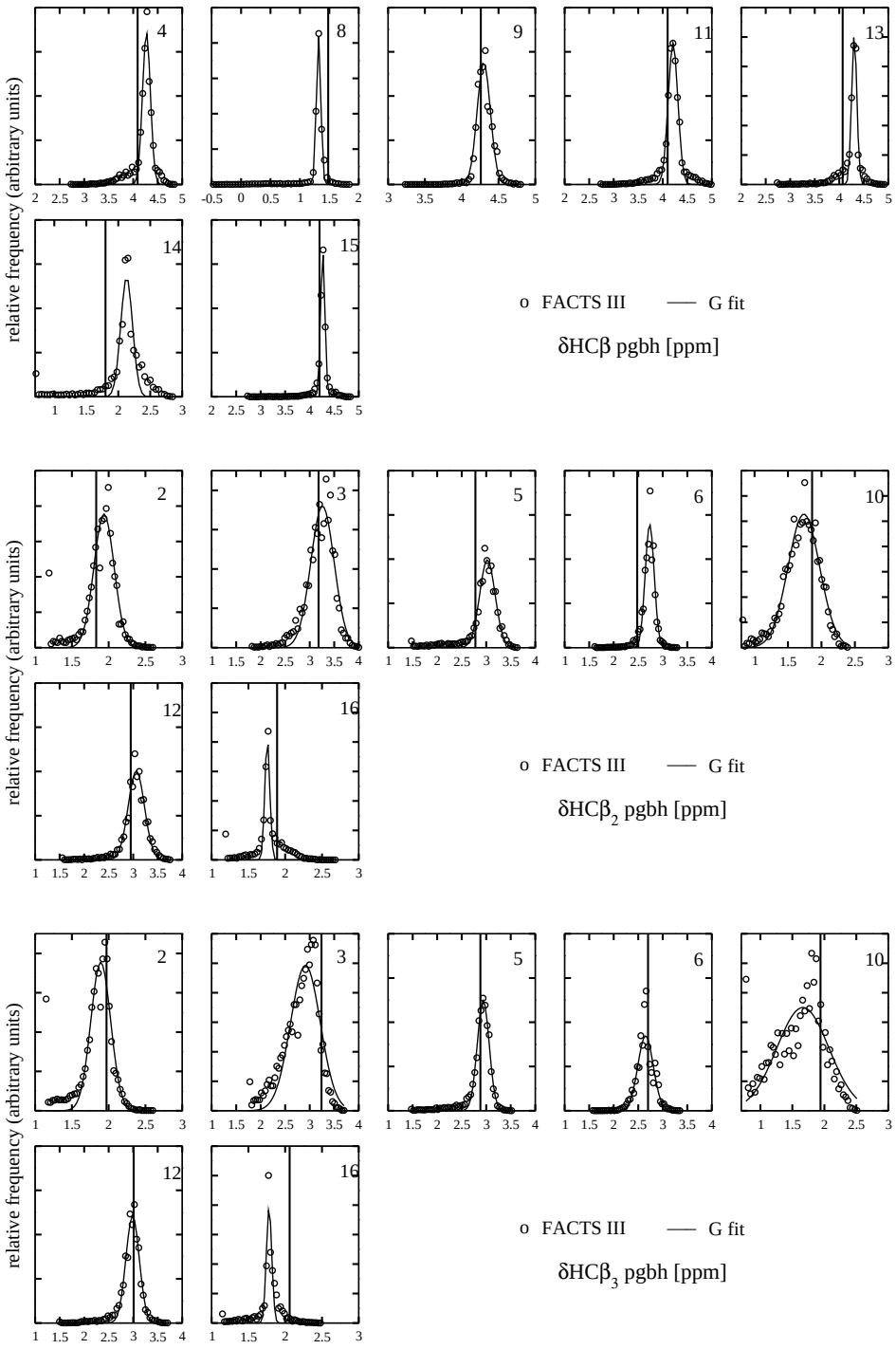


Figure 9.27: pgbh HCβ, HCβ_2 and HCβ_3 CS calculated via SHIFTX program from the FACTS simulations (4 μs) at 280 K with FACTS III. Vertical lines represent the experimental shifts. See Tab. 9.3, Tab. 9.4 and Fig. 9.24 for quantitative analysis.

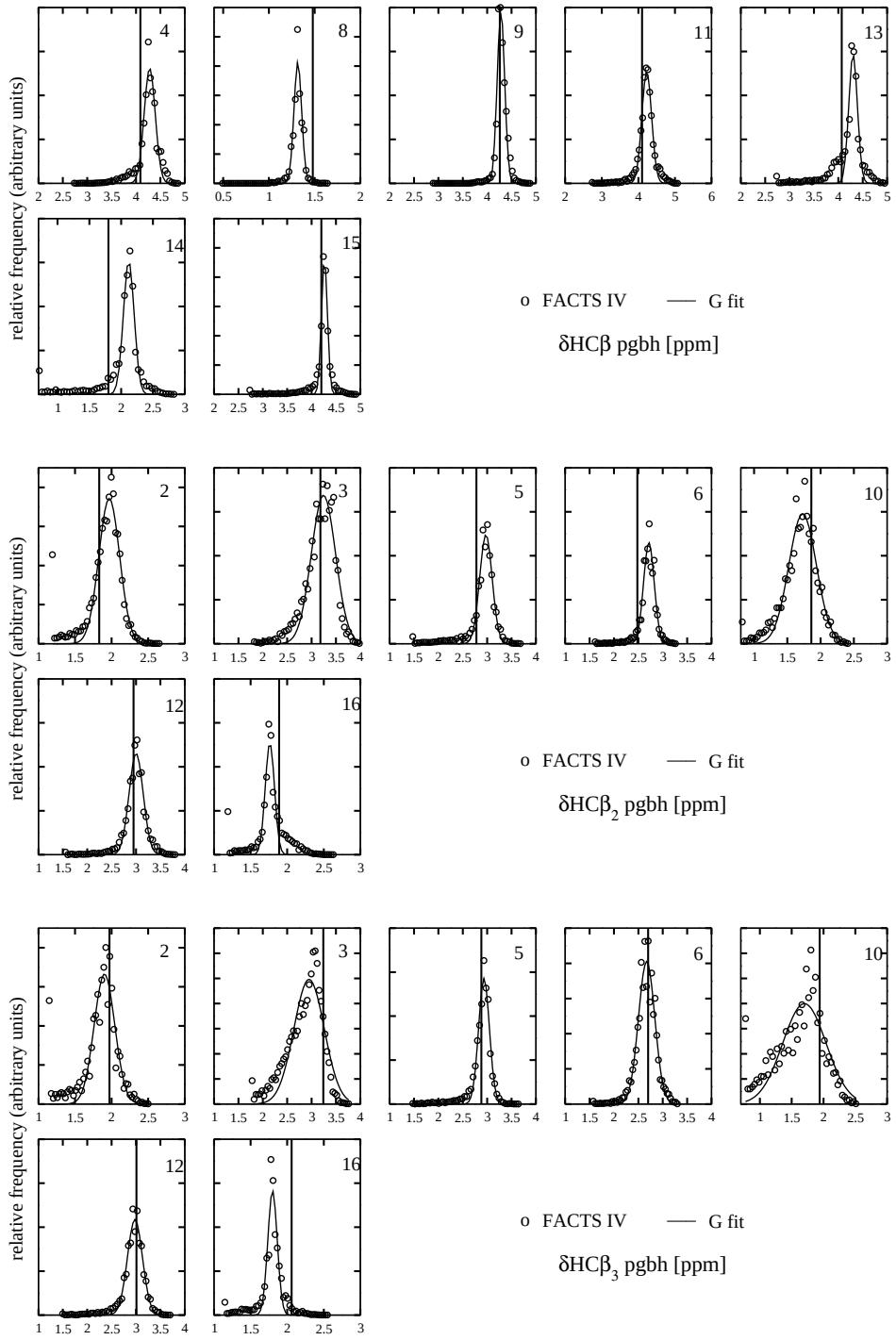


Figure 9.28: pgbh HCβ, HCβ_2 and HCβ_3 CS calculated via SHIFTX program from the FACTS simulations (4 μs) at 280 K with FACTS IV. Vertical lines represent the experimental shifts. See Tab. 9.3, Tab. 9.4 and Fig. 9.24 for quantitative analysis.

res.	exp. HCβ	FACTS sim. HCβ	I σHCβ	FACTS sim. HCβ	II σHCβ	FACTS sim. HCβ	III σHCβ	FACTS sim. HCβ	IV σHCβ
2	**1.83**	1.99	0.11	1.99	0.09	1.94	0.14	1.97	0.14
2	**1.97**	1.92	0.10	1.95	0.07	1.89	0.13	1.90	0.14
3	**3.18**	3.18	0.20	3.20	0.14	3.26	0.24	3.24	0.25
3	**3.24**	2.95	0.28	3.12	0.21	2.91	0.30	2.95	0.31
4	**4.09**	4.29	0.08	4.44	0.15	4.27	0.08	4.28	0.11
5	**2.78**	2.99	0.09	2.95	0.11	3.02	0.16	2.97	0.13
5	**2.88**	2.89	0.10	2.91	0.11	2.94	0.12	2.94	0.11
6	**2.48**	2.76	0.12	2.76	0.15	2.73	0.09	2.72	0.11
6	**2.70**	2.61	0.07	2.66	0.31	2.66	1.00	2.67	0.17
8	**1.48**	1.33	0.10	1.31	0.06	1.32	0.04	1.32	0.04
9	**4.26**	4.34	0.15	4.29	0.11	4.30	0.09	4.28	0.08
10	**1.86**	1.80	0.22	1.77	0.21	1.74	0.24	1.73	0.21
10	**1.94**	1.85	0.20	1.69	0.36	1.66	0.41	1.70	0.34
11	**4.10**	4.18	0.18	4.18	0.17	4.21	0.10	4.23	0.13
12	**2.95**	3.02	0.10	3.01	0.12	3.06	0.17	3.00	0.14
12	**3.01**	2.97	0.12	2.92	0.13	2.99	0.14	2.98	0.15
15	**4.20**	4.26	0.03	4.24	0.05	4.27	0.04	4.26	0.06

Table 9.5: Comparison between experimental (bold) and simulated δHCβ CS of pgbh with FACTS (at 280 K).

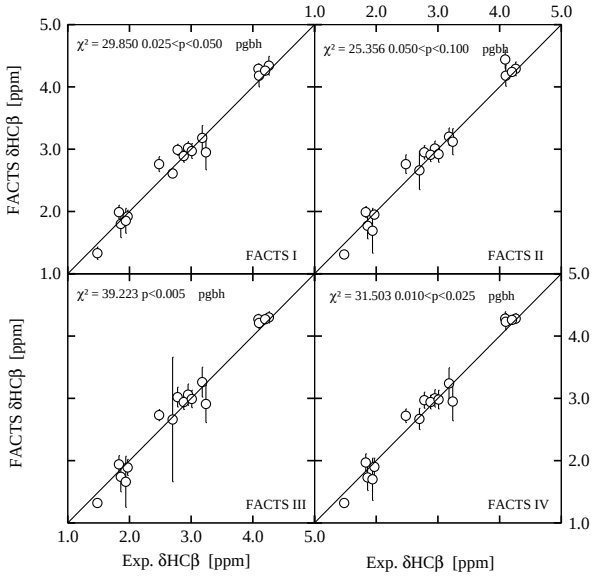


Figure 9.29: HCβ CS of FACTS simulations with pgbh in comparison with experimental values (280 K).

par.	DF	χ^2	p
FACTS I	17	29.8503	0.025¡p¡0.050
FACTS II	17	25.3562	0.050¡p¡0.100
FACTS III	17	39.2228	p¡0.005
FACTS IV	17	31.5028	0.010¡p¡0.025

Table 9.6: Statistical analysis of pgbh δ HCβ shifts (at 280 K) for experimental and calculated values.

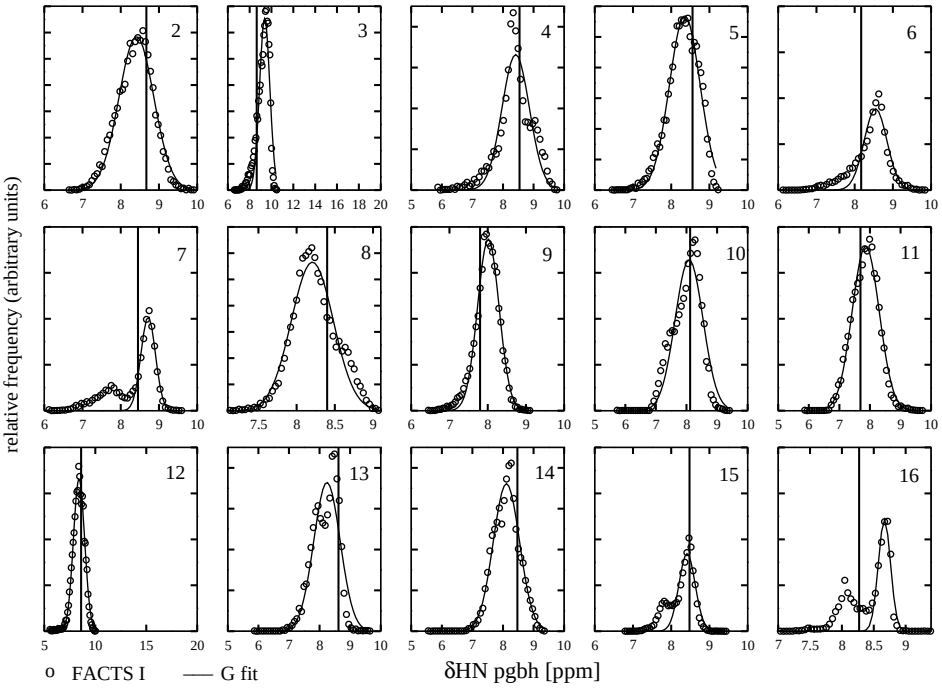


Figure 9.30: pgbh HN CS calculated via SHIFTX program from the FACTS simulations (4 μs) at 280 K with FACTS I. Vertical lines represent the experimental shifts. See Tab. 9.7, Tab. 9.8 and Fig. 9.34 for quantitative analysis.

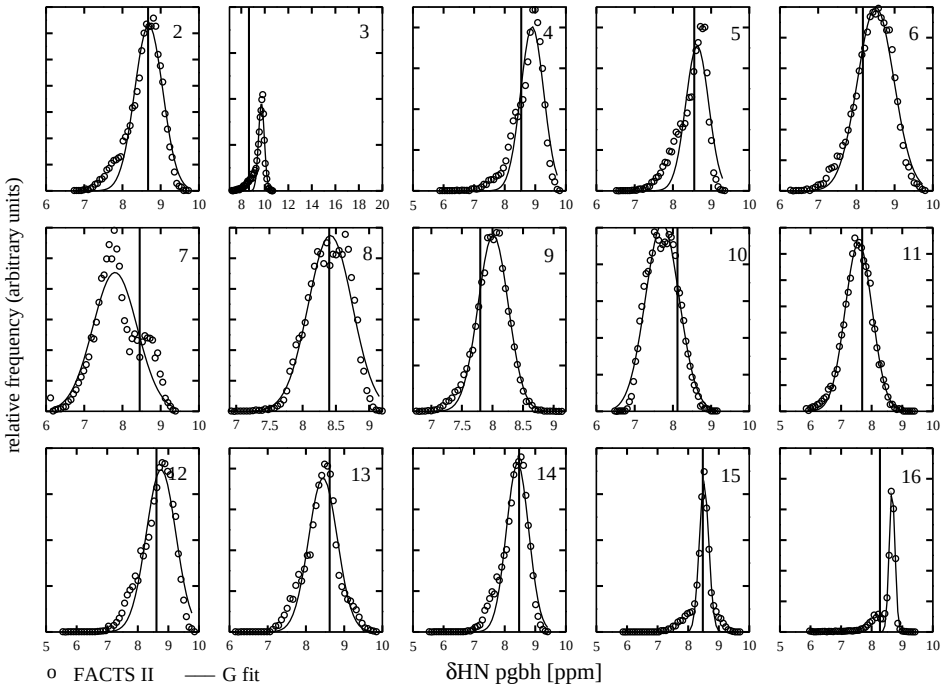


Figure 9.31: pgbh HN CS calculated via SHIFTX program from the FACTS simulations (4 μs) at 280 K with FACTS II. Vertical lines represent the experimental shifts. See Tab. 9.7, Tab. 9.8 and Fig. 9.34 for quantitative analysis.

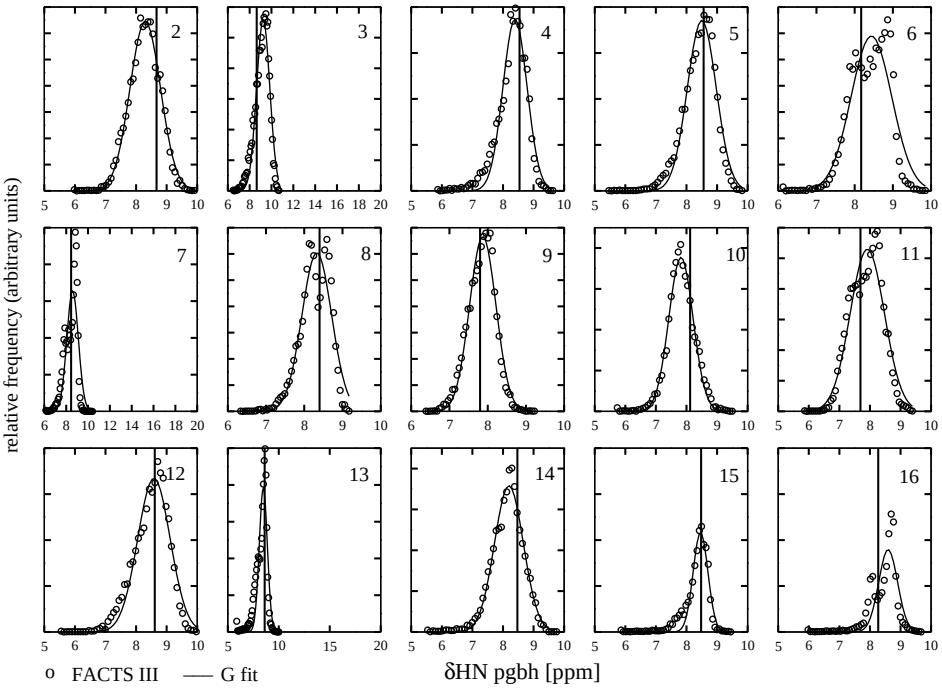


Figure 9.32: pgbh HN CS calculated via SHIFTX program from the FACTS simulations (4 μs) at 280 K with FACTS III. Vertical lines represent the experimental shifts. See Tab. 9.7, Tab. 9.8 and Fig. 9.34 for quantitative analysis.

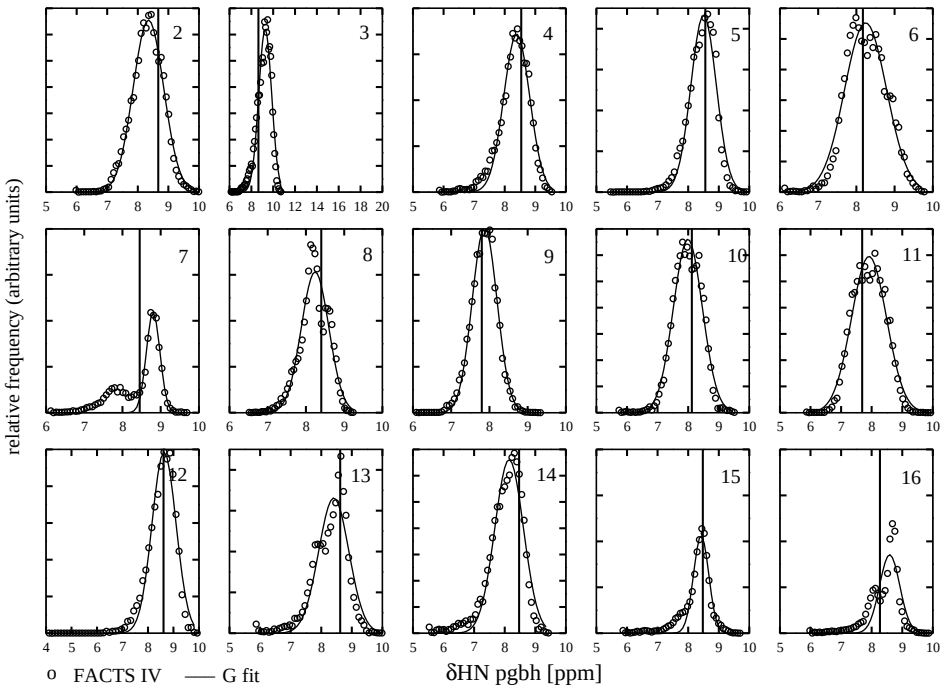


Figure 9.33: pgbh HN CS calculated via SHIFTX program from the FACTS simulations (4 μs) at 280 K with FACTS IV. Vertical lines represent the experimental shifts. See Tab. 9.7, Tab. 9.8 and Fig. 9.34 for quantitative analysis.

res.	exp. HN	FACTS sim. HN	I σHN	FACTS sim. HN	II σHN	FACTS sim. HN	III σHN	FACTS sim. HN	IV σHN
2	**8.67**	8.41	0.46	8.69	0.37	8.30	0.50	8.35	0.50
3	**8.65**	9.38	0.47	9.71	0.25	9.24	0.60	9.27	0.59
4	**8.54**	8.42	0.45	8.90	0.38	8.39	0.39	8.39	0.42
5	**8.56**	8.36	0.40	8.64	0.30	8.50	0.47	8.50	0.41
6	**8.17**	8.55	0.30	8.51	0.47	8.44	0.53	8.24	0.55
7	**8.45**	8.72	0.19	7.80	0.58	8.59	0.53	8.80	0.18
8	**8.40**	8.21	0.27	8.42	0.33	8.33	0.39	8.24	0.36
9	**7.80**	8.02	0.28	8.01	0.25	7.87	0.35	7.87	0.33
10	**8.12**	8.09	0.46	7.72	0.44	7.82	0.40	7.99	0.47
11	**7.69**	7.86	0.44	7.58	0.44	7.91	0.56	7.91	0.57
12	**8.61**	8.40	0.55	8.76	0.48	8.58	0.53	8.63	0.47
13	**8.62**	8.24	0.43	8.45	0.38	8.48	0.40	8.41	0.49
14	**8.47**	8.11	0.43	8.42	0.36	8.19	0.47	8.14	0.47
15	**8.48**	8.42	0.21	8.52	0.16	8.46	0.24	8.42	0.25
16	**8.27**	8.67	0.10	8.67	0.09	8.59	0.31	8.59	0.33

Table 9.7: Comparison between experimental (bold) and simulated δHN CS of pgbh with FACTS (at 280 K).

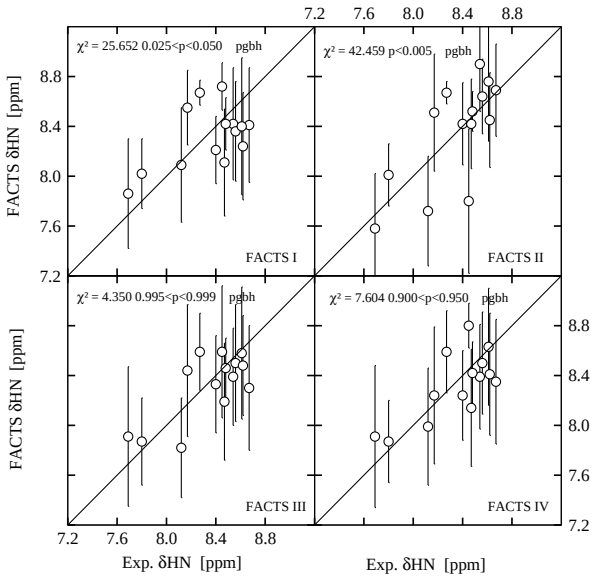


Figure 9.34: HN CS of FACTS simulations with pgbh in comparison with experimental values (280 K).

par.	DF	χ^2	p
FACTS I	15	25.6522	0.025¡p¡0.050
FACTS II	15	42.4593	p¡0.005
FACTS III	15	4.35024	0.995¡p¡0.999
FACTS IV	15	7.60429	0.900¡p¡0.950

Table 9.8: Statistical analysis of pgbh δ HN shifts (at 280 K) for experimental and calculated values.

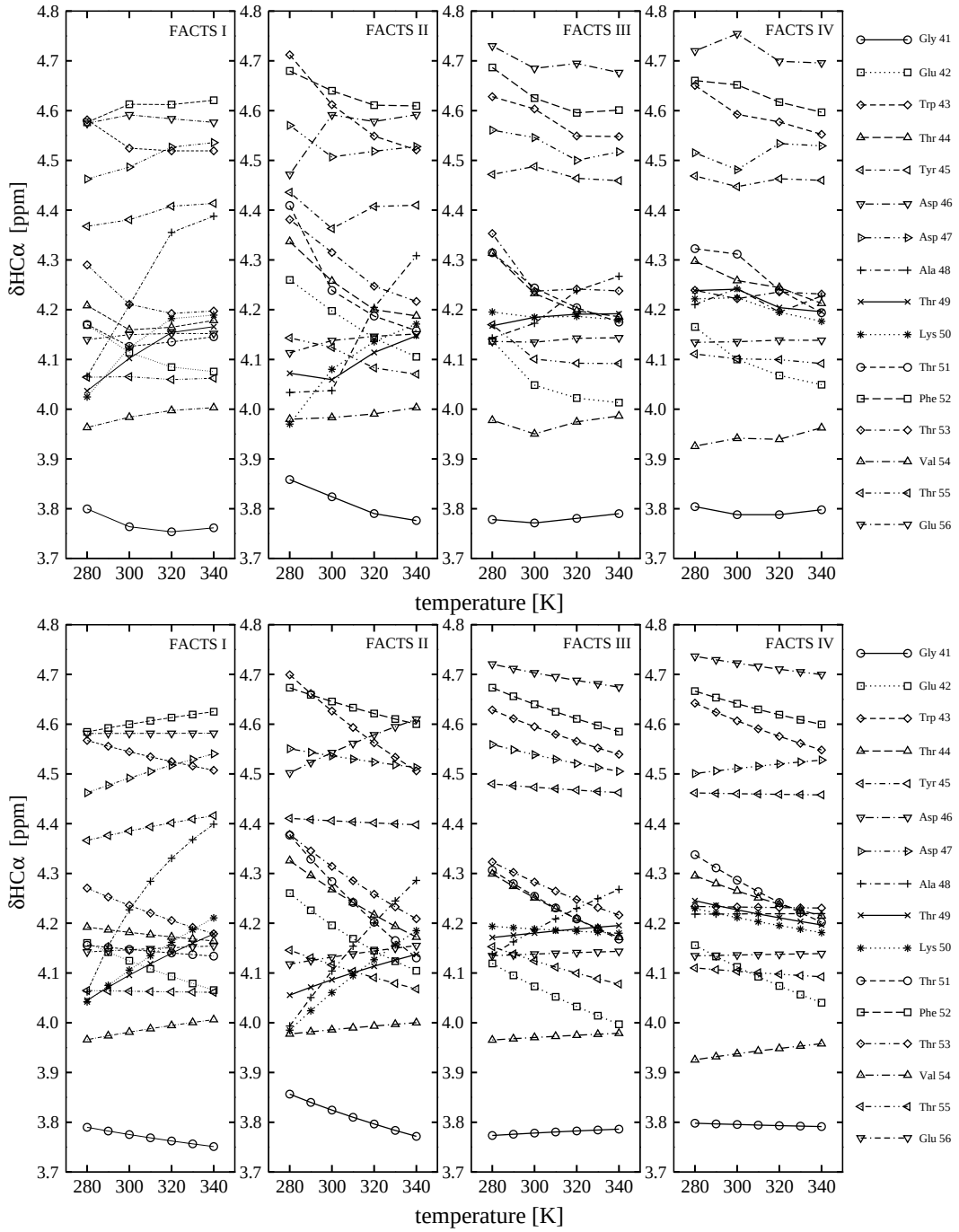


Figure 9.35: Simulated (upper panel) and fitted trends (bottom panel) of pgbh δHCα shifts as a function of temperature with FACTS, to be compared with the (NMR) experimental melting curves of Cα protons of pgbh peptide in 99.996% ^2H_2O with 5 %mM sodium phosphate buffer (p^{2}H 7.0) shown in Fig. 3 of [43]. See Tab. 9.9 for quantitative analysis.

	$T_m^{exp.}$	$\Delta H_m^{exp.}$	$T_m^{F.\,I}$	$\Delta H_m^{F.\,I}$	$T_m^{F.\,II}$	$\Delta H_m^{F.\,II}$	$T_m^{F.\,III}$	$\Delta H_m^{F.\,III}$	$T_m^{F.\,IV}$	$\Delta H_m^{F.\,IV}$
res.	[K]	[kJ/mol]	[K]	[kJ/mol]	[K]	[kJ/mol]	[K]	[kJ/mol]	[K]	[kJ/mol]
Glu 42	**297**	**50.8**	296.4	18.7	76.5	505.4	286.7	28.3	275.8	40.1
Trp 43	**288**	**50.8**	296.7	13.0	289.9	32.9	296.8	17.8	296.6	18.5
Thr 44	**291**	**50.9**	297.0	6.5	287.6	31.3	285.6	30.1	297.0	16.3
Tyr 45	**304**	**44.2**	297.0	-11.0	297.0	3.1	297.1	4.1	297.0	1.0
Asp 47	**296**	**48.6**	296.9	-16.1	296.9	8.6	296.9	11.8	297.6	-6.3
Asp 48	**296**	**62.4**	139.1	442	316.9	-37.9	297.5	-23.2	297.0	-0.3
Ala 49	**291**	**61.1**	283.4	-33.9	297.7	-16.4	297.0	-5.7	297.0	10.8
Thr 50	**295**	**52.8**	291.5	-29.8	288.8	-33.8	297.0	3.3	297.0	10.7
Lys 51	**291**	**46.9**	297.0	5.1	282.3	40.7	276.7	40.1	299.5	23.1
Thr 52	**290**	**53.2**	297.0	-9.2	296.7	15.3	296.2	17.9	297.1	14.1
Phe 53	**290**	**60.6**	295.8	18.5	275.3	47.4	294.5	21.0	297.0	0.7
Thr 56	**285**	**52.6**	297.0	-2.9	297.0	-8.5	297.0	-2.1	297.0	-1.1
ave.	293	53	282.1	-126.9	275.2	49.0	293.3	12.0	295.5	10.6

Table 9.9: Comparison of experimental and calculate values from MD simulations with FACTS (F.) of the transition temperature T_m and enthalpy ΔH_m of each Cα carbon, fitted over HCα thermal dependence.

Acetyl-(AAQAA)$_3$-amide

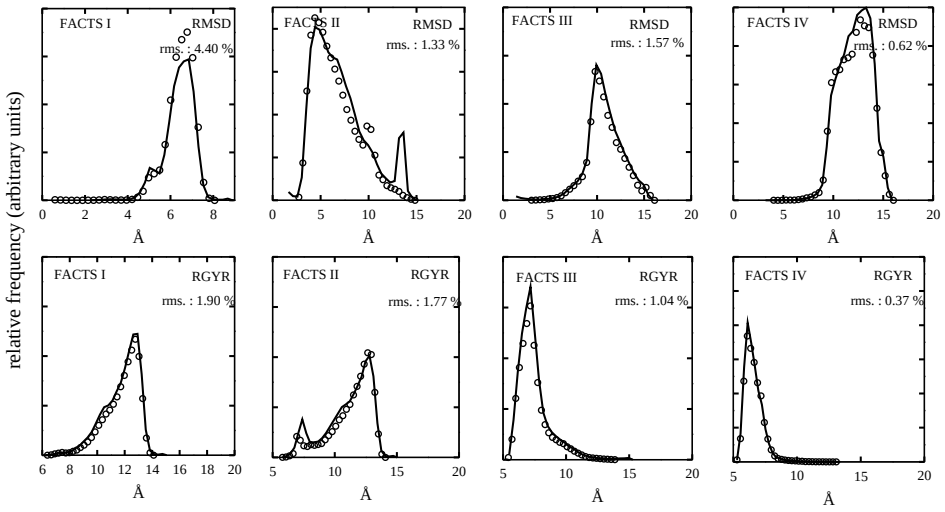


Figure 9.36: Convergence test based on halving the 4 μs long act2 simulations with FACTS (at 274 K) in two sections. The reference structure for the RMSD timeseries is a perfect alpha helix. Circles refer to the first section of the trajectory; solid lines refer to the second half (the % deviation between the RMSD (and RGYR) distributions in the first and in the second half is shown).

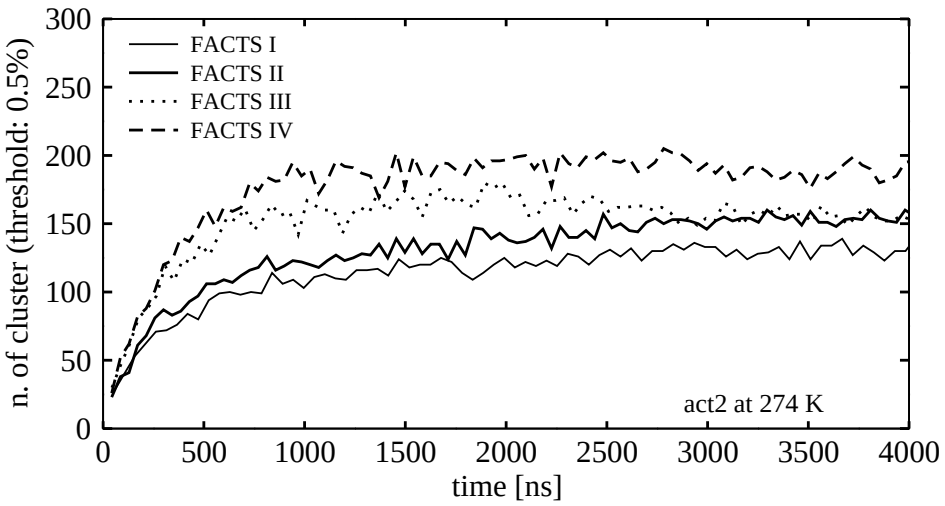
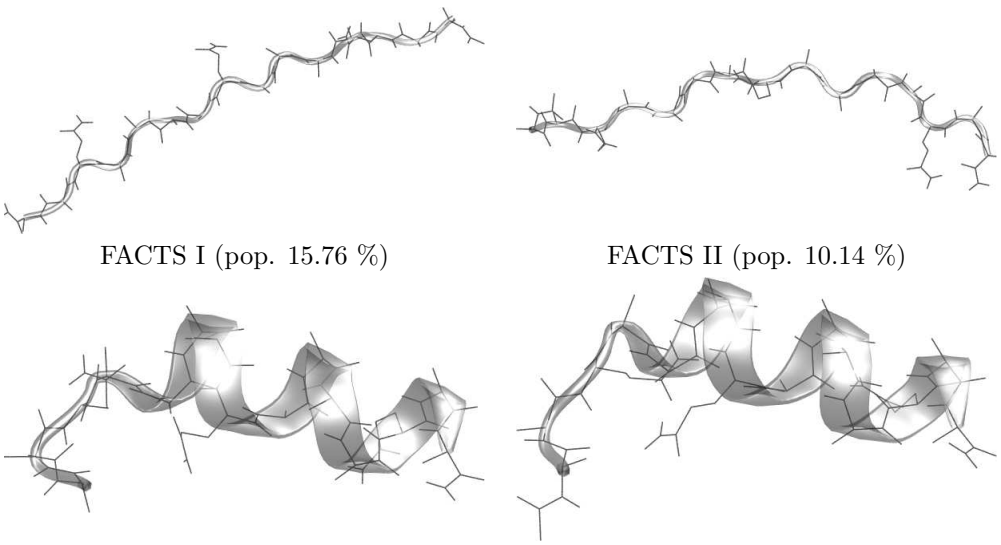


Figure 9.37: Convergence test for the same simulations as in Fig. 9.36 based on the number of significantly populated clusters.



FACTS I (pop. 15.76 %) FACTS II (pop. 10.14 %)

FACTS III (pop. 8.69 %) FACTS IV (pop. 7.50 %)

Figure 9.38: Central conformations of peptide act2 (at 274 K). The RMSD-clustering was performed with Wordom with a cutoff of 2.5 . Simulations are 4 μs long.

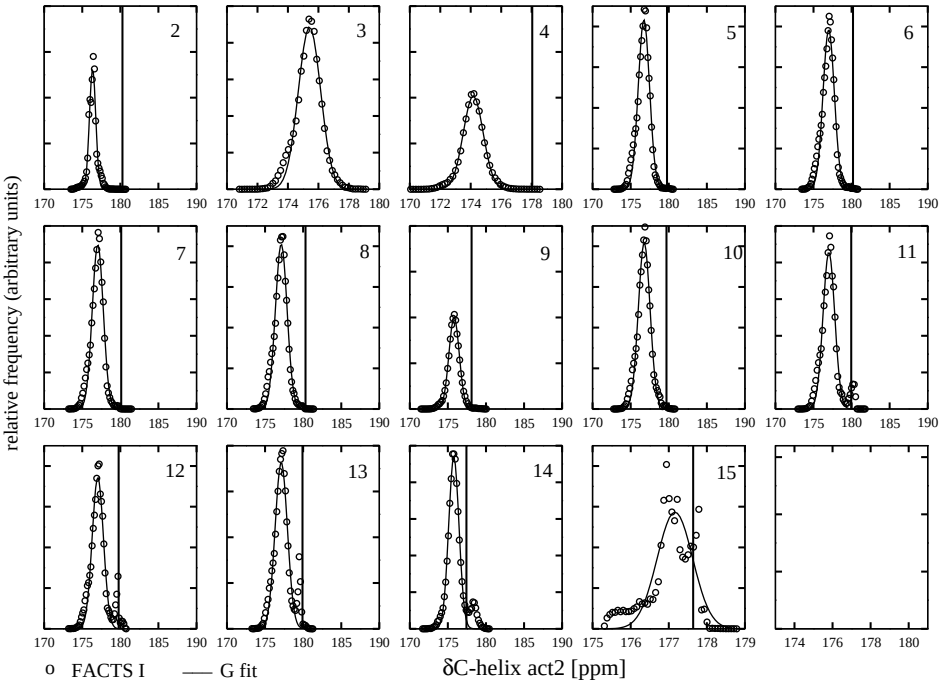


Figure 9.39: act2 δC shifts calculated via SHIFTX program from the FACTS simulations (6 μs) with FACTS I in comparison with Sholongo helix shifts (274 K).

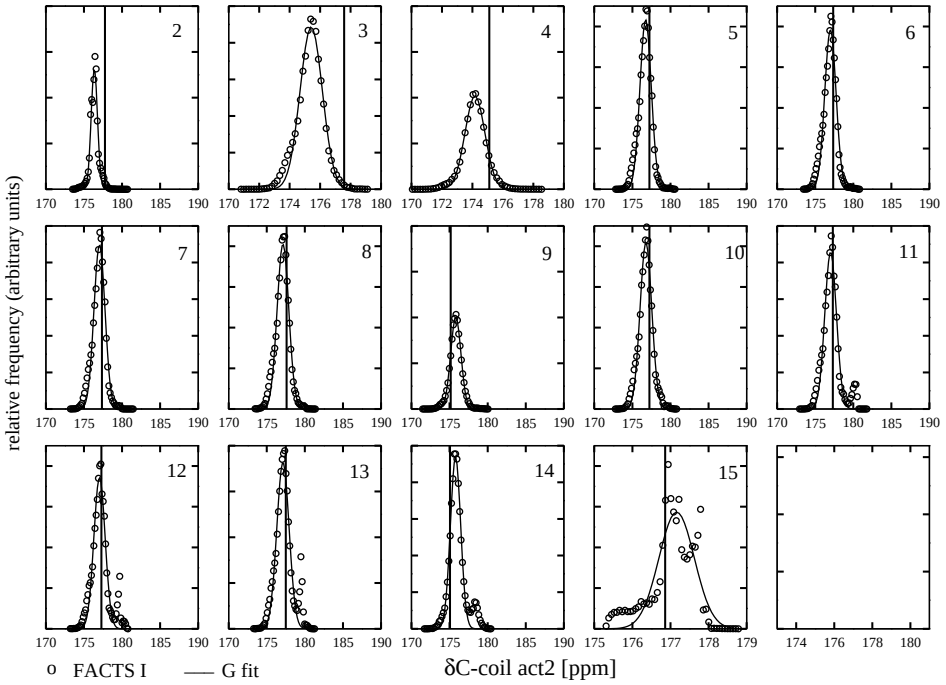


Figure 9.40: act2 δC shifts calculated via SHIFTX program from the FACTS simulations (6 μs) with FACTS I in comparison with Sholongo coil shifts (274 K).

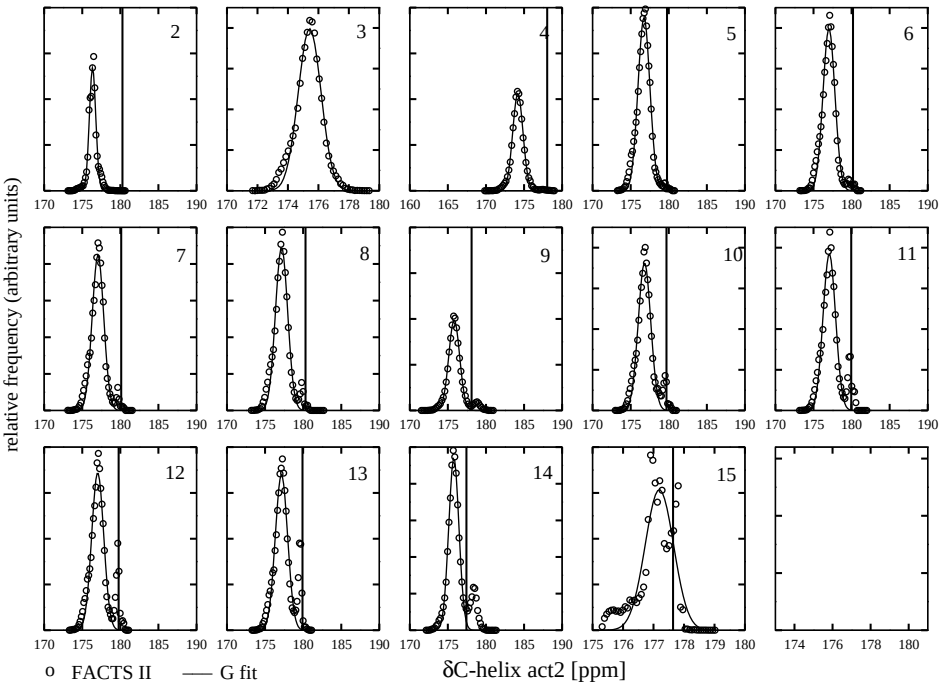


Figure 9.41: act2 δC shifts calculated via SHIFTX program from the FACTS simulations (6 μs) with FACTS II in comparison with Sholongo helix shifts.

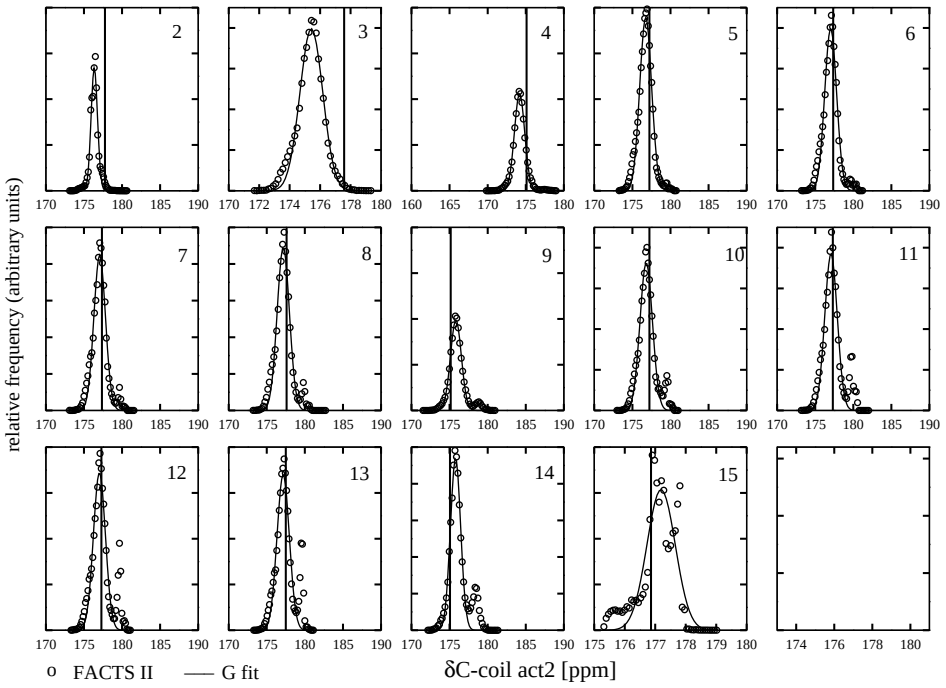


Figure 9.42: act2 δC shifts calculated via SHIFTX program from the FACTS simulations (6 μs) with FACTS II in comparison with Sholongo coil shifts.

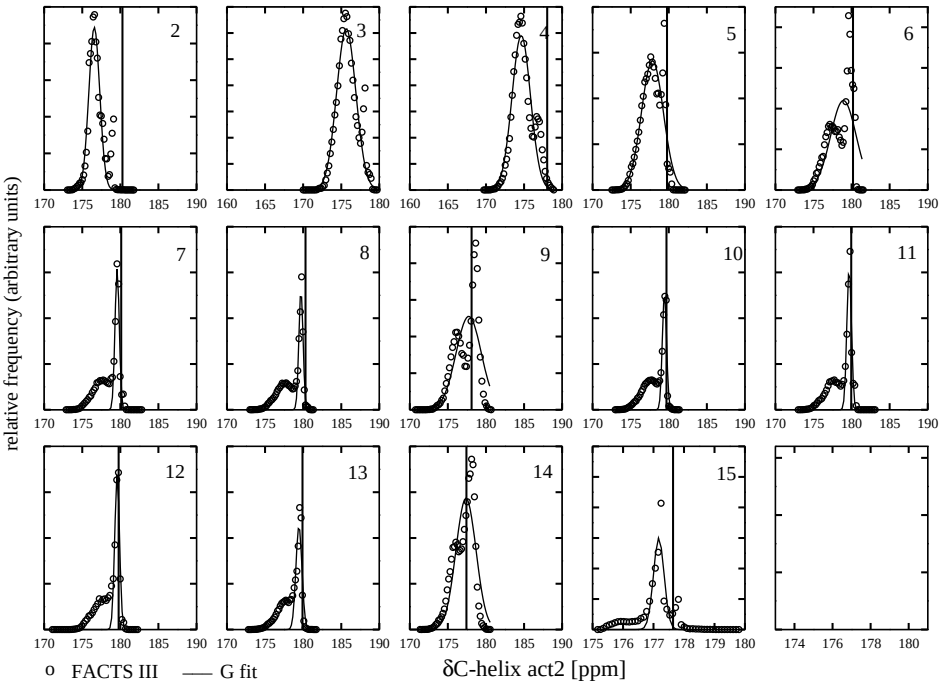


Figure 9.43: act2 δC shifts calculated via SHIFTX program from the FACTS simulations (6 μs) with FACTS III in comparison with Sholongo helix coil shifts.

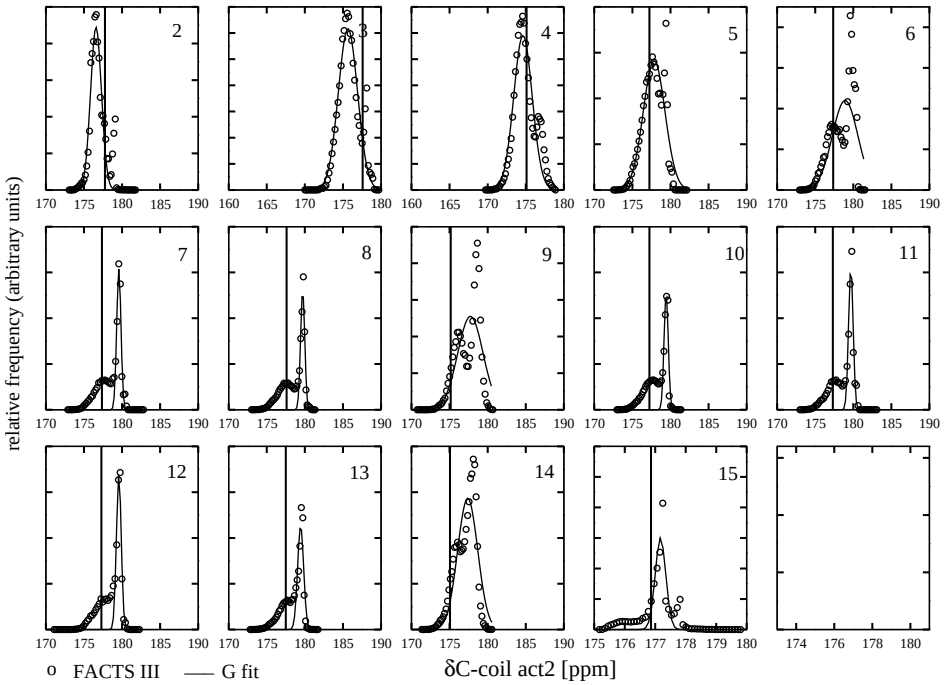


Figure 9.44: act2 δC shifts calculated via SHIFTX program from the FACTS simulations (6 μs) with FACTS III in comparison with Sholongo coil shifts.

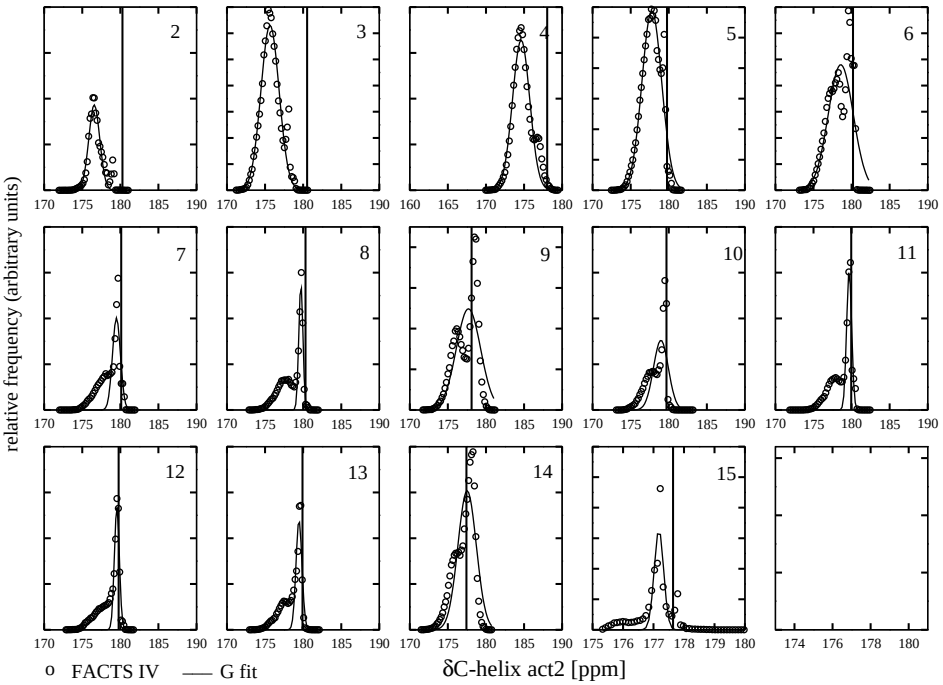


Figure 9.45: act2 δC shifts calculated via SHIFTX program from the FACTS simulations (6 μs) with FACTS III in comparison with Sholongo coil shifts.

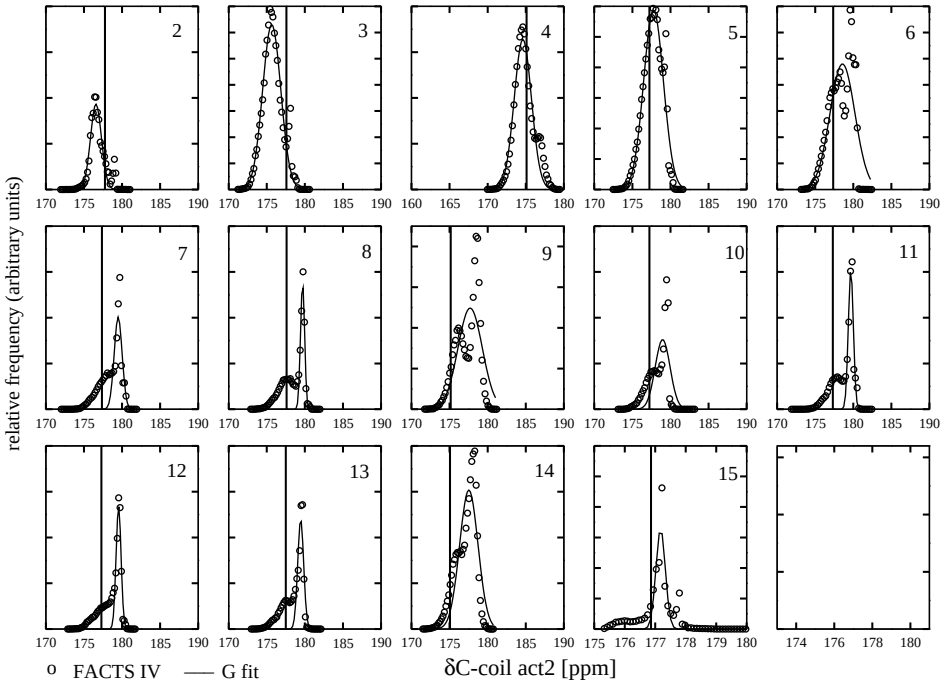


Figure 9.46: act2 δC shifts calculated via SHIFTX program from the FACTS simulations (6 μs) with FACTS IV in comparison with Sholongo helix shifts (bottom).

res.	exp. C	FACTS sim. C	I σC	FACTS sim. C	II σC	FACTS sim. C	III σC	FACTS sim. C	IV σC
2	**180.28**	176.36	0.42	176.35	0.42	176.58	0.71	176.57	0.72
3	**180.52**	175.42	0.73	175.43	0.75	175.66	1.26	175.65	1.19
4	**178.04**	174.17	0.65	174.19	0.67	174.63	1.22	174.62	1.13
5	**179.77**	176.74	0.74	176.75	0.78	177.85	1.43	177.78	1.30
6	**180.19**	177.00	0.74	177.03	0.78	178.93	1.68	178.58	1.64
7	**180.11**	177.03	0.72	177.06	0.76	179.58	0.28	179.49	0.52
8	**180.31**	177.12	0.77	177.17	0.80	179.71	0.27	179.70	0.27
9	**178.11**	175.81	0.67	175.83	0.71	177.73	1.66	177.70	1.60
10	**179.7**	176.78	0.77	176.81	0.82	179.43	0.28	178.97	0.98
11	**179.96**	177.02	0.76	177.08	0.82	179.68	0.30	179.68	0.33
12	**179.77**	177.04	0.74	177.03	0.79	179.58	0.31	179.54	0.33
13	**179.91**	177.12	0.81	177.15	0.81	179.44	0.40	179.45	0.39
14	**177.43**	175.78	0.67	175.78	0.67	177.38	1.32	177.53	1.21
15	**177.64**	177.16	0.47	177.19	0.47	177.17	0.17	177.18	0.15

Table 9.10: Comparison between experimental (bold) and simulated δC-helix shifts of act2 with FACTS.

res.	exp. C	FACTS sim. C	I σC	FACTS sim. C	II σC	FACTS sim. C	III σC	FACTS sim. C	IV σC
2	**177.75**	176.36	0.42	176.35	0.42	176.58	0.71	176.57	0.72
3	**177.58**	175.42	0.73	175.43	0.75	175.66	1.26	175.65	1.19
4	**175.11**	174.17	0.65	174.19	0.67	174.63	1.22	174.62	1.13
5	**177.23**	176.74	0.74	176.75	0.78	177.85	1.43	177.78	1.30
6	**177.36**	177.00	0.74	177.03	0.78	178.93	1.68	178.58	1.64
7	**177.36**	177.03	0.72	177.06	0.76	179.58	0.28	179.49	0.52
8	**177.60**	177.12	0.77	177.17	0.80	179.71	0.27	179.70	0.27
9	**175.15**	175.81	0.67	175.83	0.71	177.73	1.66	177.70	1.60
10	**177.24**	176.78	0.77	176.81	0.82	179.43	0.28	178.97	0.98
11	**177.32**	177.02	0.76	177.08	0.82	179.68	0.30	179.68	0.33
12	**177.29**	177.04	0.74	177.03	0.79	179.58	0.31	179.54	0.33
13	**177.49**	177.12	0.81	177.15	0.81	179.44	0.40	179.45	0.39
14	**175.05**	175.78	0.67	175.78	0.67	177.38	1.32	177.53	1.21
15	**176.86**	177.16	0.47	177.19	0.47	177.17	0.17	177.18	0.15

Table 9.11: Comparison between experimental (in bold) and simulated δC-coil shifts of act2 with FACTS.

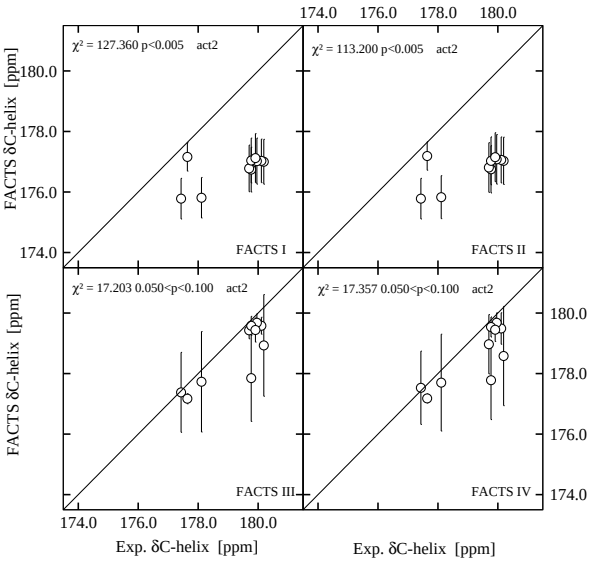


Figure 9.47: δC CS of FACTS simulations with act2 in comparison with Sholongo's (helix) experimental values.

par.	DF	χ^2	p
FACTS I	10	127.36	p¡0.005
FACTS II	10	113.2	p¡0.005
FACTS III	10	17.2028	0.050¡p¡0.100
FACTS IV	10	17.3573	0.050¡p¡0.100

Table 9.12: Statistical analysis of act2 δC-helix shifts between experimental and calculated values.

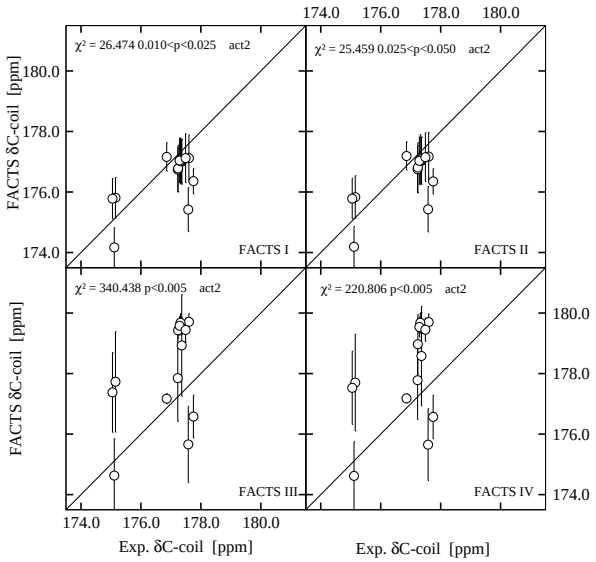


Figure 9.48: δC CS of FACTS simulations with pgbh in comparison with Sholongo's (coil) experimental values.

par.	DF	χ^2	p
FACTS I	14	26.4736	0.010¡p¡0.025
FACTS II	14	25.4593	0.025¡p¡0.050
FACTS III	14	340.438	p¡0.005
FACTS IV	14	220.806	p¡0.005

Table 9.13: Statistical analysis of act2 δC-coil shifts between experimental and calculated values.

gsgs peptide

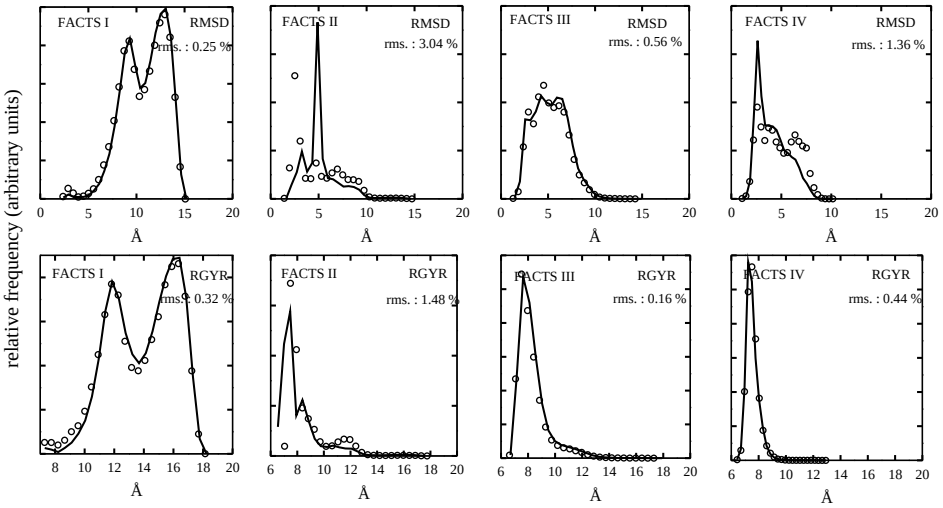


Figure 9.49: Convergence test based on halving the 6 μs long gsgs simulations with FACTS (at 300 K) in two sections. The reference structure for the RMSD timeseries is the SASA native state. Circles refer to the first section of the trajectory; solid lines refer to the second half (the % deviation between the RMSD (and RGYR) distributions in the first and in the second half is shown).

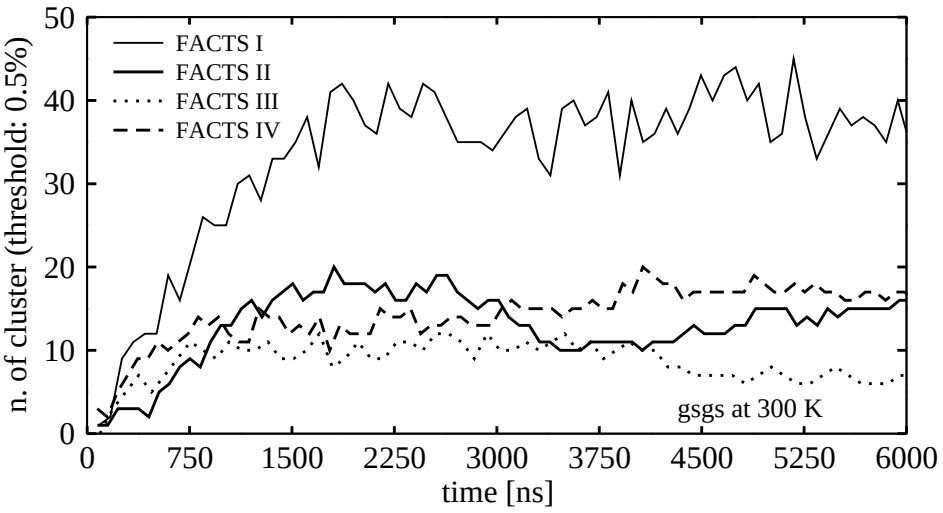
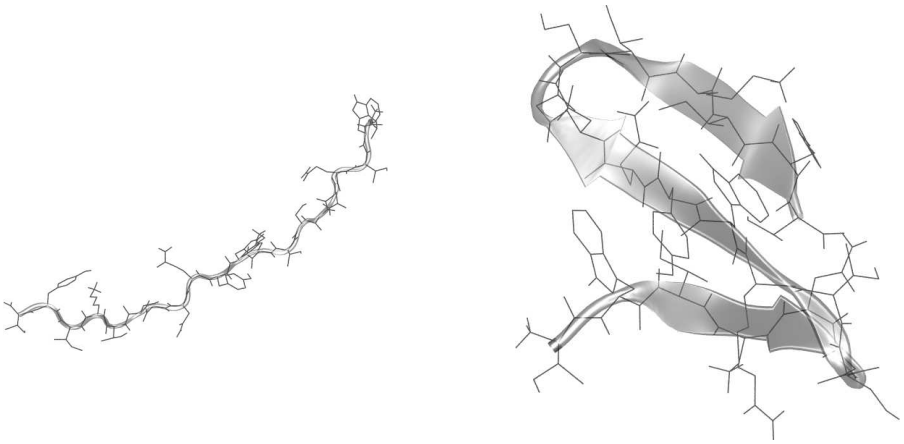
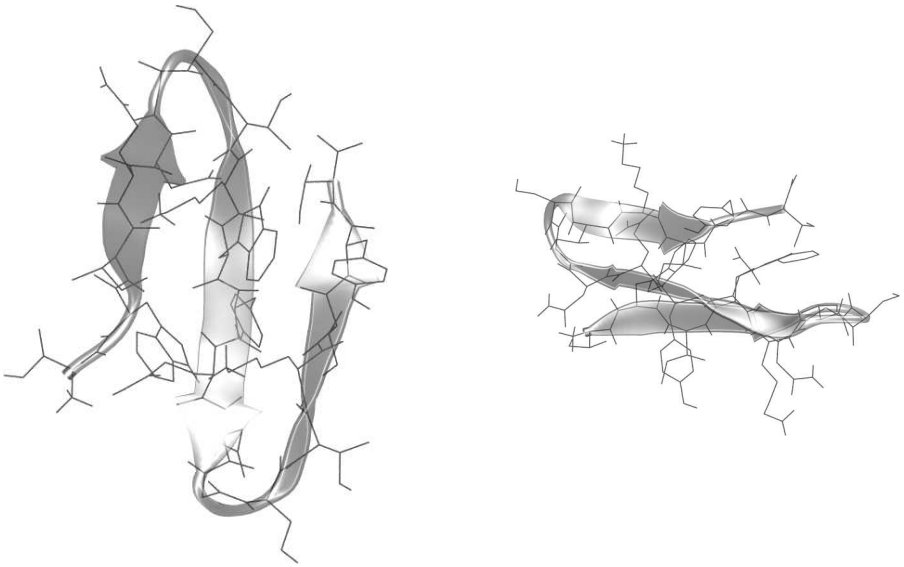


Figure 9.50: Convergence test for the same simulations as in Fig. 9.49 based on the number of significantly populated clusters.



FACTS I (pop. 2.13 %) FACTS II (pop. 19.52 %)



FACTS III (pop. 10.47 %) FACTS IV (pop. 30.83 %)

Figure 9.51: Central conformations of peptide gsgs (at 300 K). The RMSD-clustering was performed with Wordom with a cutoff of 2.5 . Simulations are 6 μs long.

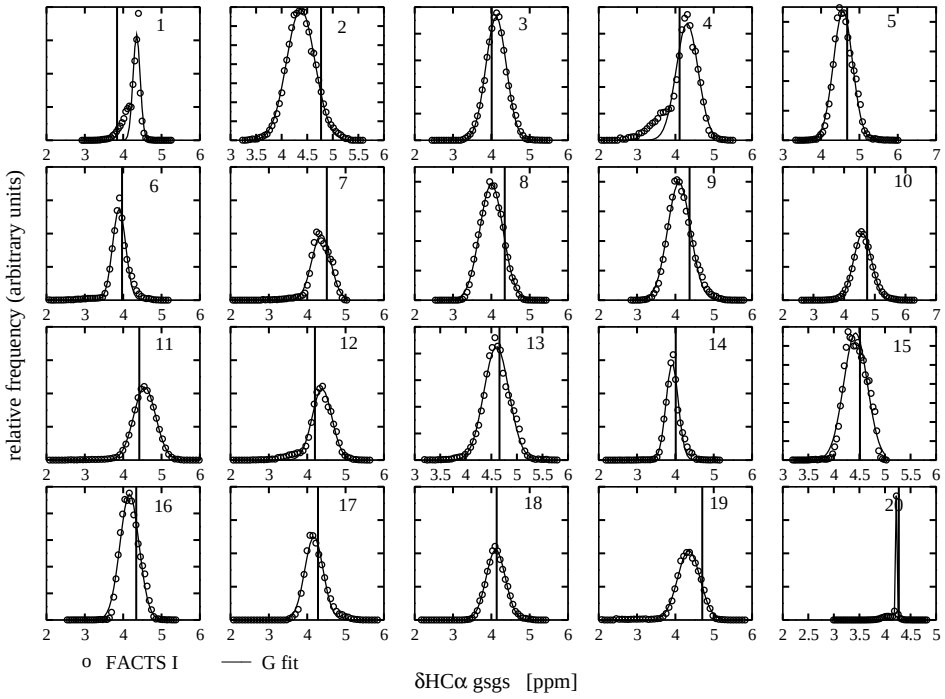


Figure 9.52: gsgs HCα CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS I. Vertical lines represent the experimental shifts. See Tab. 9.14, Tab. 9.15 and Fig. 9.56 for quantitative analysis.

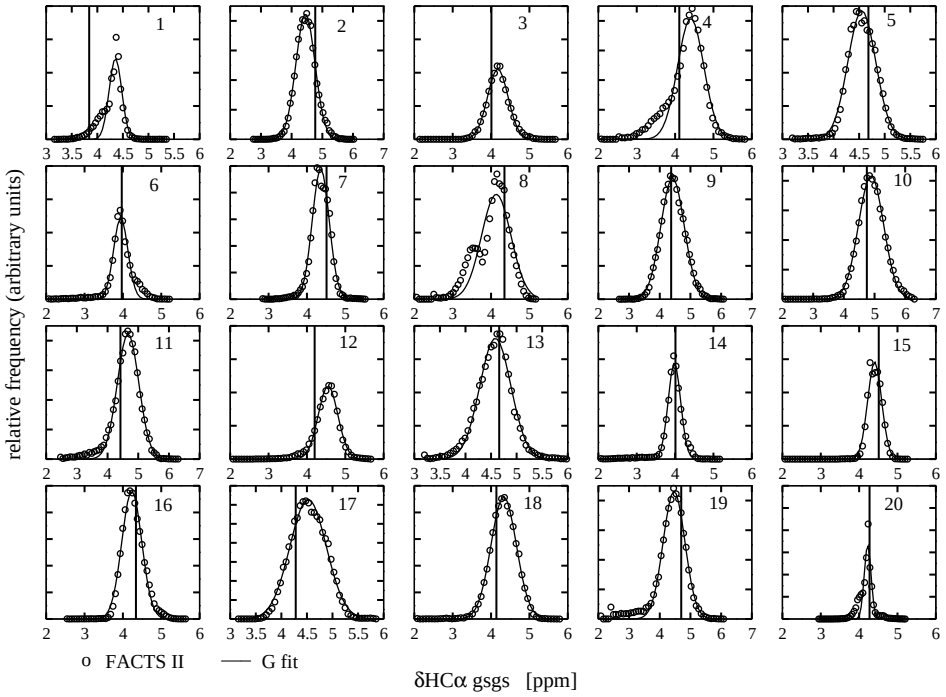


Figure 9.53: gsgs HCα CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS II. Vertical lines represent the experimental shifts. See Tab. 9.14, Tab. 9.15 and Fig. 9.56 for quantitative analysis.

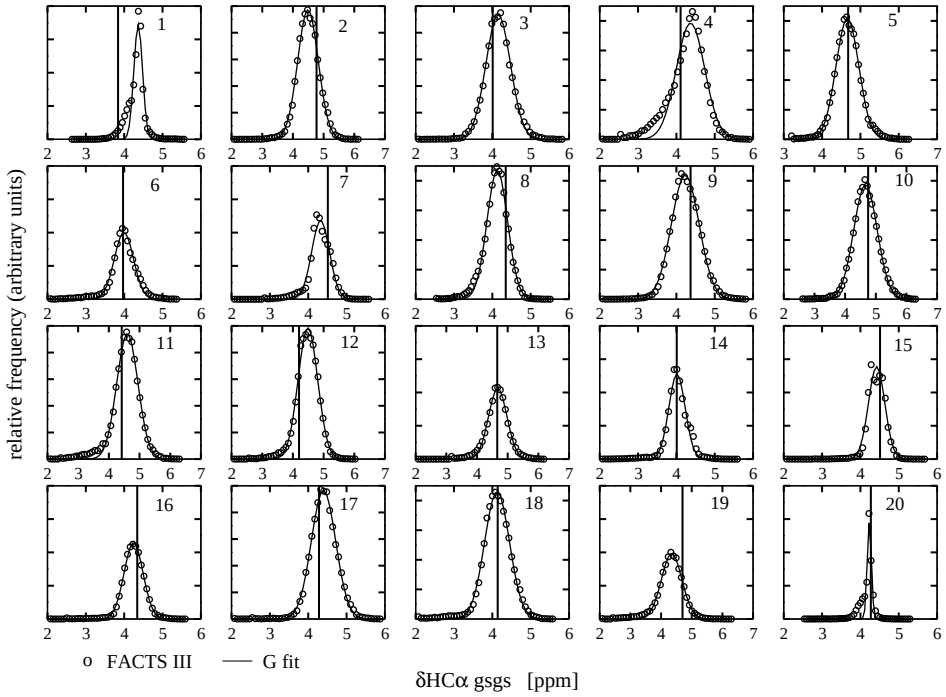


Figure 9.54: gsgs HCα CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS III. Vertical lines represent the experimental shifts. See Tab. 9.14, Tab. 9.15 and Fig. 9.56 for quantitative analysis.

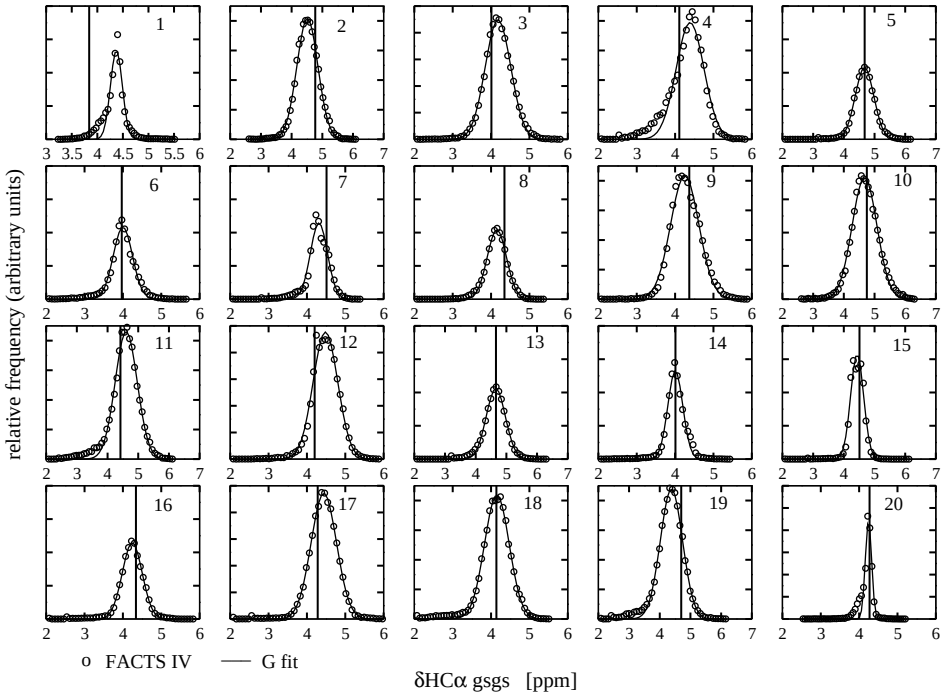


Figure 9.55: gsgs HCα CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS IV. Vertical lines represent the experimental shifts. See Tab. 9.14, Tab. 9.15 and Fig. 9.56 for quantitative analysis.

res.	exp. HCα	FACTS sim. HCα	I σHCα	FACTS sim. HCα	II σHCα	FACTS sim. HCα	III σHCα	FACTS sim. HCα	IV σHCα
2	**4.77**	4.37	0.28	4.45	0.31	4.49	0.34	4.51	0.34
3	**4.01**	4.12	0.27	4.18	0.27	4.15	0.32	4.18	0.33
4	**4.11**	4.31	0.29	4.41	0.34	4.37	0.37	4.39	0.36
5	**4.68**	4.57	0.25	4.55	0.27	4.66	0.30	4.67	0.30
6	**3.97**	3.89	0.17	3.94	0.19	3.99	0.25	4.00	0.25
7	**4.51**	4.36	0.24	4.36	0.22	4.32	0.23	4.31	0.22
8	**4.35**	4.00	0.30	4.14	0.38	4.13	0.27	4.15	0.25
9	**4.37**	4.09	0.30	4.42	0.38	4.23	0.36	4.25	0.37
10	**4.75**	4.60	0.29	4.89	0.41	4.66	0.38	4.68	0.40
11	**4.42**	4.56	0.29	4.67	0.35	4.60	0.34	4.61	0.36
12	**4.20**	4.38	0.26	4.55	0.26	4.49	0.32	4.48	0.32
13	**4.66**	4.60	0.24	4.58	0.30	4.68	0.30	4.66	0.31
14	**4.01**	3.92	0.16	3.98	0.16	4.02	0.21	4.01	0.20
15	**4.51**	4.42	0.23	4.40	0.18	4.42	0.21	4.43	0.21
16	**4.34**	4.16	0.25	4.23	0.26	4.25	0.25	4.24	0.26
17	**4.28**	4.17	0.24	4.52	0.36	4.40	0.32	4.44	0.33
18	**4.14**	4.11	0.25	4.33	0.33	4.12	0.32	4.15	0.32
19	**4.70**	4.35	0.29	4.47	0.35	4.35	0.33	4.38	0.36
20	**4.27**	4.22	0.02	4.22	0.09	4.23	0.06	4.24	0.08

Table 9.14: Comparison between experimental (bold) and simulated δHCα CS of gsgs with FACTS (at 300 K).

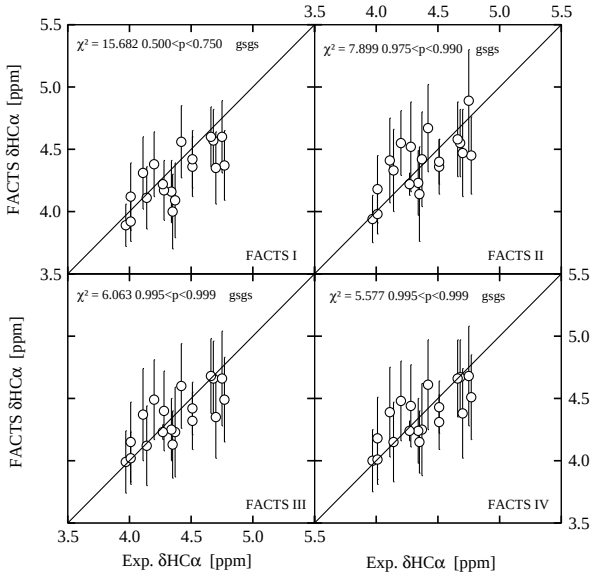


Figure 9.56: HCα CS of FACTS simulations with gsgs in comparison with experimental values (300 K).

par.	DF	χ^2	p
FACTS I	19	15.6817	0.500¡p¡0.750
FACTS II	19	7.89895	0.975¡p¡0.990
FACTS III	19	6.06333	0.995¡p¡0.999
FACTS IV	19	5.57689	0.995¡p¡0.999

Table 9.15: Statistical analysis of gsgs δ HCα shifts (at 300 K) for experimental and calculated values.

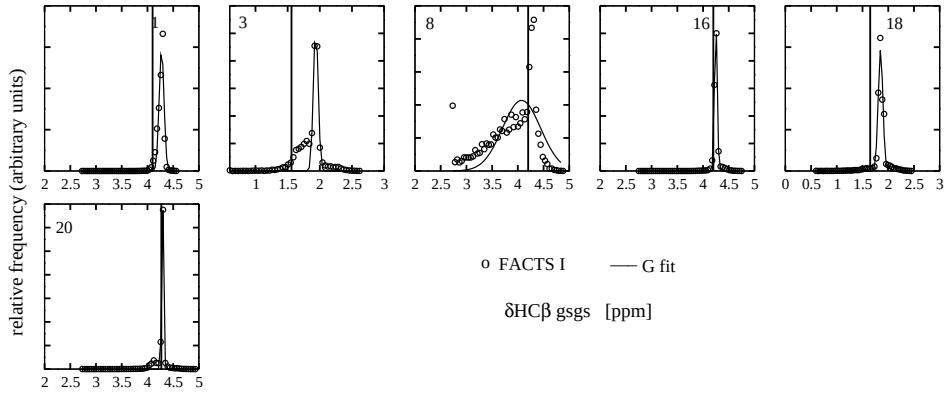


Figure 9.57: gsgs HCβ CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS I. Vertical lines represent the experimental shifts. See Tab. 9.16, Tab. 9.17 and Fig. 9.69 for quantitative analysis.

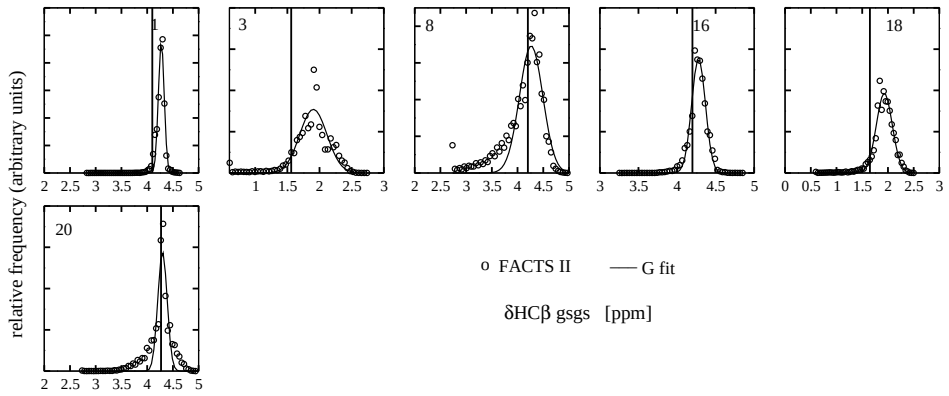


Figure 9.58: gsgs HCβ CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS II. Vertical lines represent the experimental shifts. See Tab. 9.16, Tab. 9.17 and Fig. 9.69 for quantitative analysis.

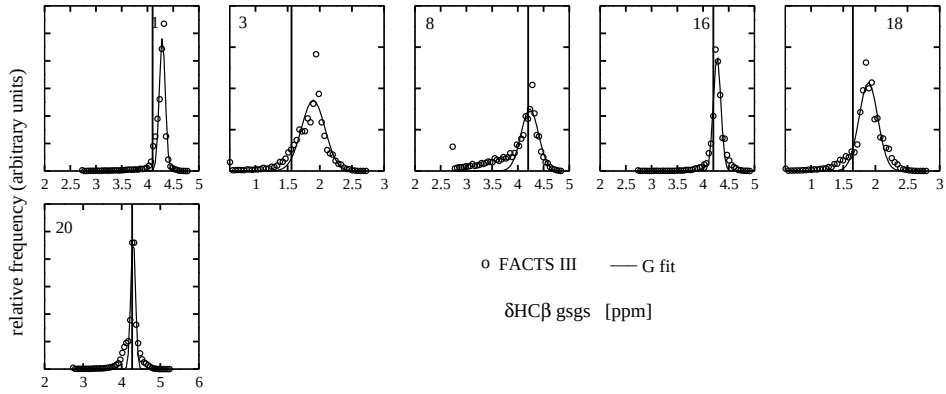


Figure 9.59: gsgs HCβ CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS III. Vertical lines represent the experimental shifts. See Tab. 9.16, Tab. 9.17 and Fig. 9.69 for quantitative analysis.

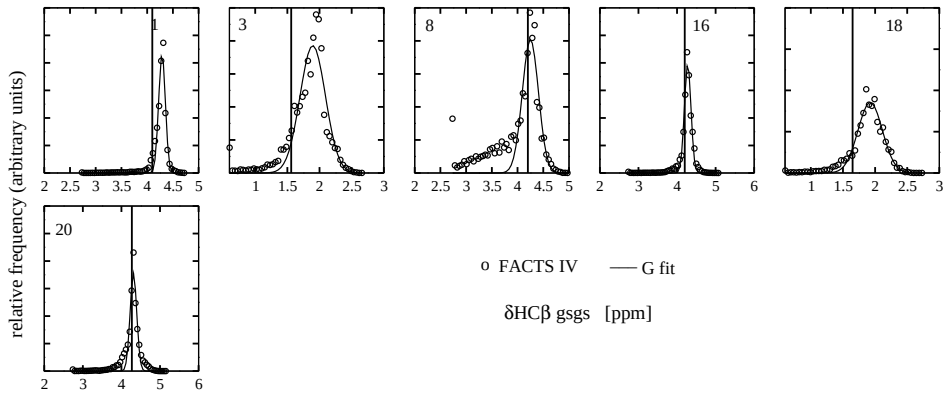


Figure 9.60: gsgs HCβ CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS IV. Vertical lines represent the experimental shifts. See Tab. 9.16, Tab. 9.17 and Fig. 9.69 for quantitative analysis.

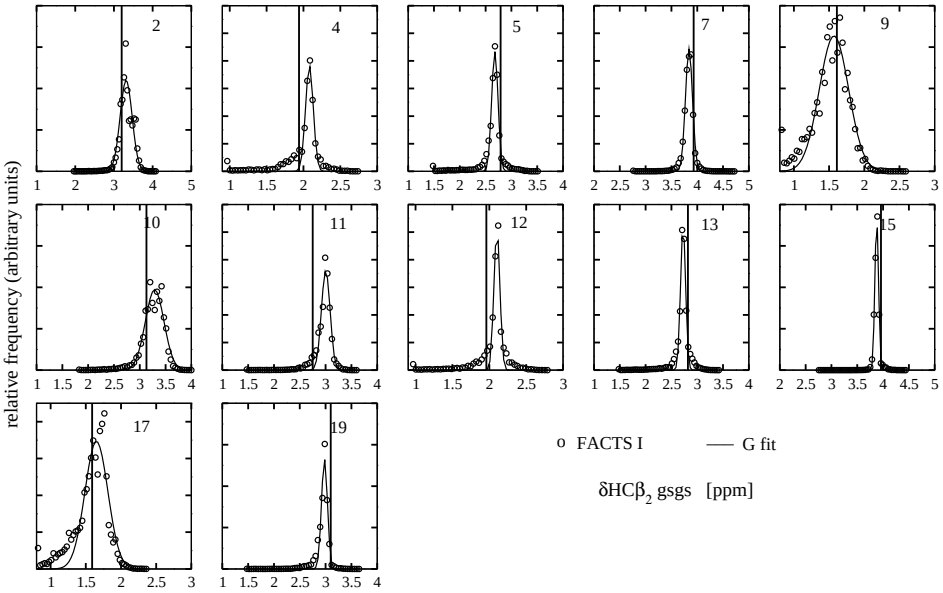


Figure 9.61: gsgs HCβ_2 CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS I. Vertical lines represent the experimental shifts. See Tab. 9.16, Tab. 9.17 and Fig. 9.69 for quantitative analysis.

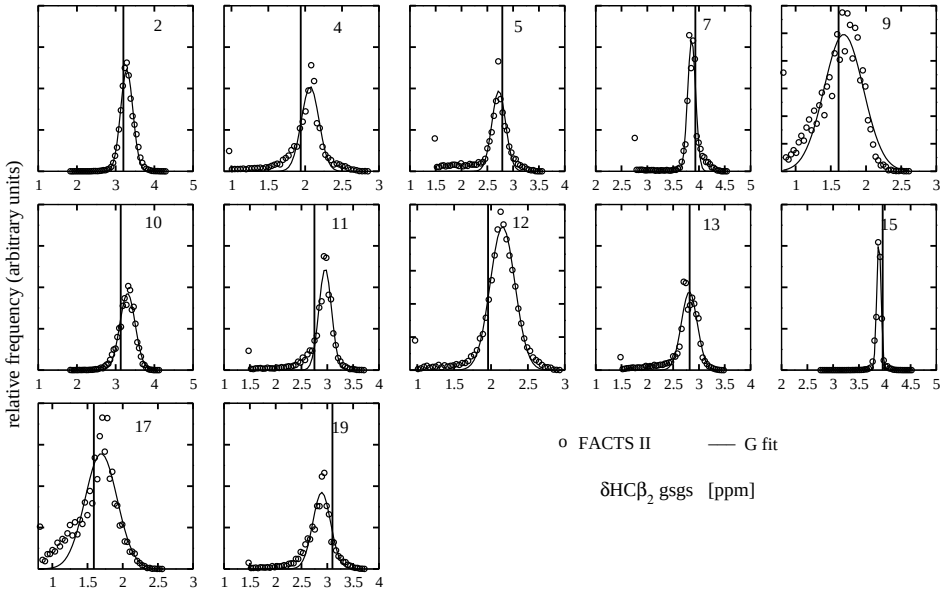


Figure 9.62: gsgs HCβ_2 CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS II. Vertical lines represent the experimental shifts. See Tab. 9.16, Tab. 9.17 and Fig. 9.69 for quantitative analysis.

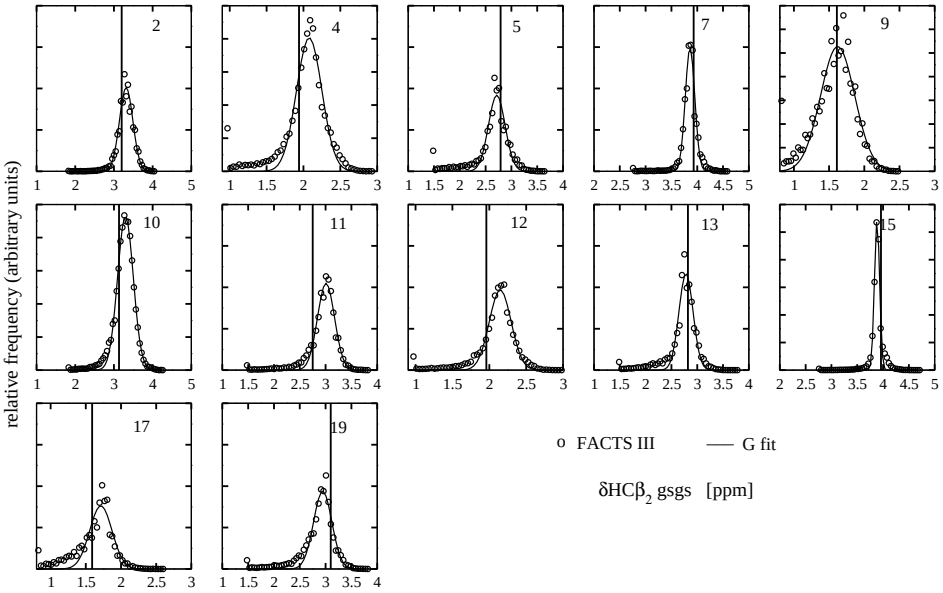


Figure 9.63: gsgs HCβ_2 CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS III. Vertical lines represent the experimental shifts. See Tab. 9.16, Tab. 9.17 and Fig. 9.69 for quantitative analysis.

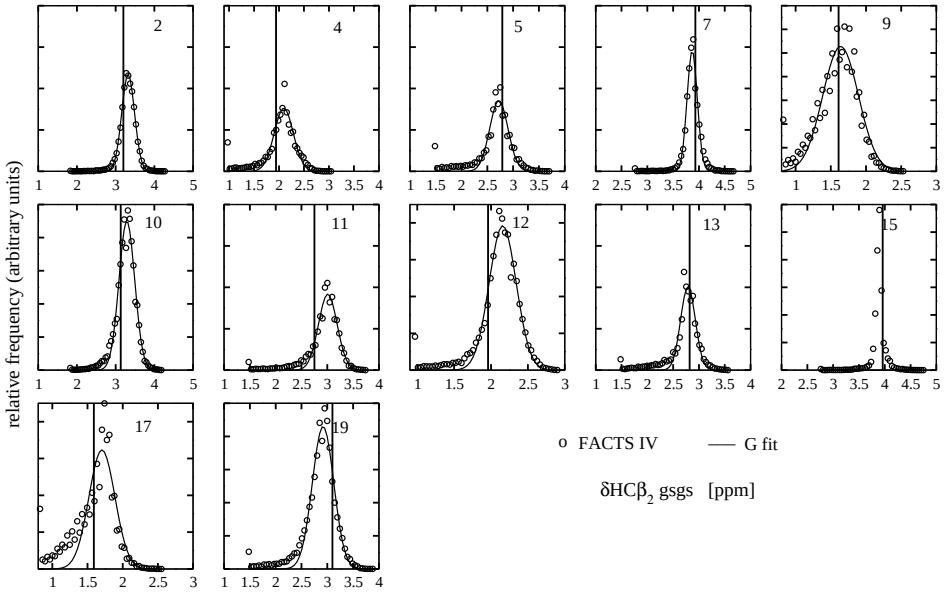


Figure 9.64: gsgs HCβ_2 CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS IV. Vertical lines represent the experimental shifts. See Tab. 9.16, Tab. 9.17 and Fig. 9.69 for quantitative analysis.

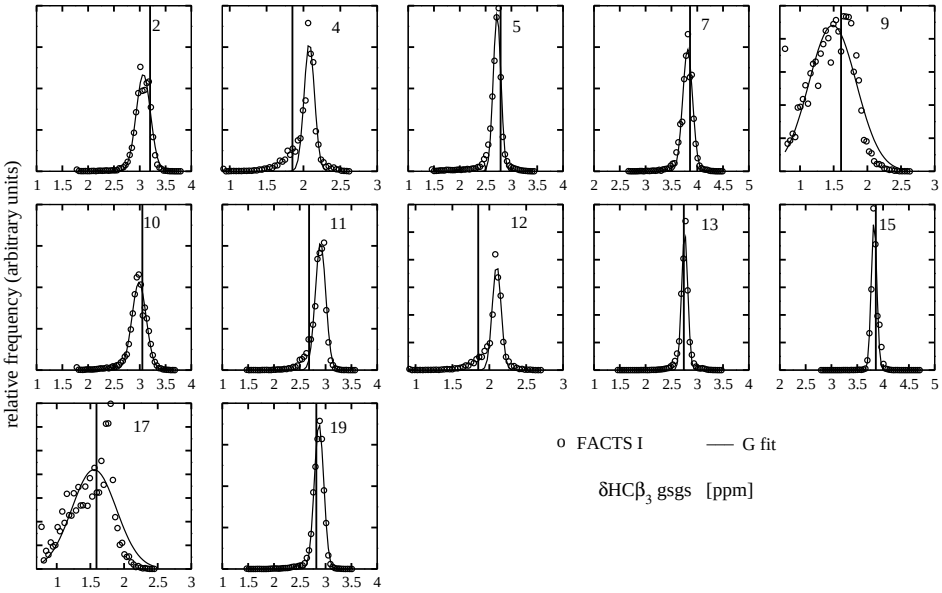


Figure 9.65: gsgs HCβ_3 CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS I. Vertical lines represent the experimental shifts. See Tab. 9.16, Tab. 9.17 and Fig. 9.69 for quantitative analysis.

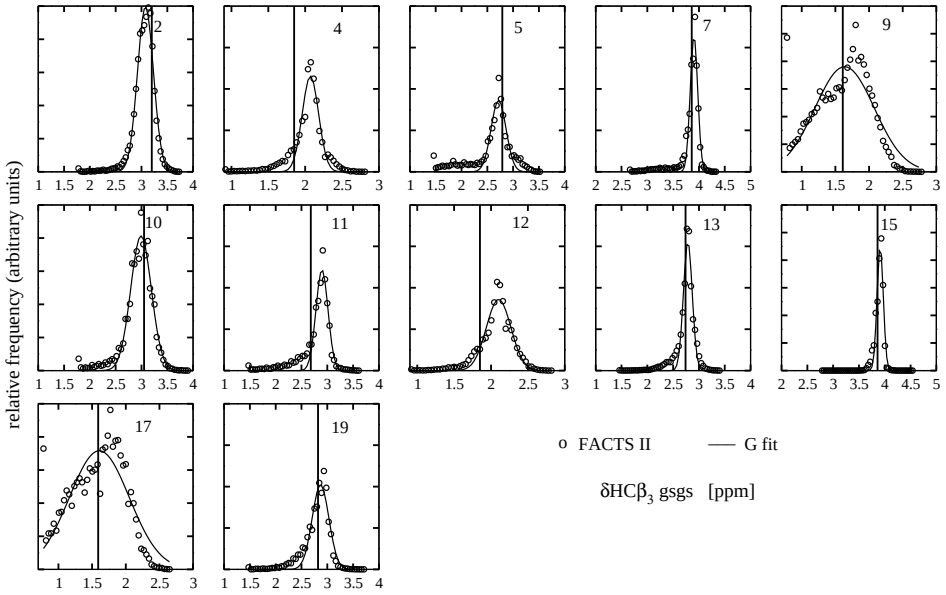


Figure 9.66: gsgs HCβ_3 CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS II. Vertical lines represent the experimental shifts. See Tab. 9.16, Tab. 9.17 and Fig. 9.69 for quantitative analysis.

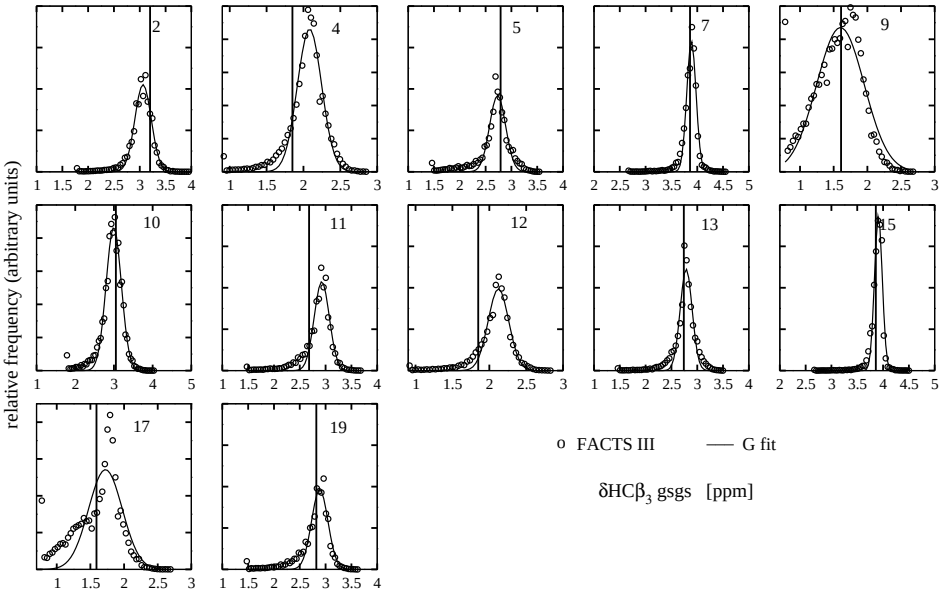


Figure 9.67: gsgs HCβ_3 CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS III. Vertical lines represent the experimental shifts. See Tab. 9.16, Tab. 9.17 and Fig. 9.69 for quantitative analysis.

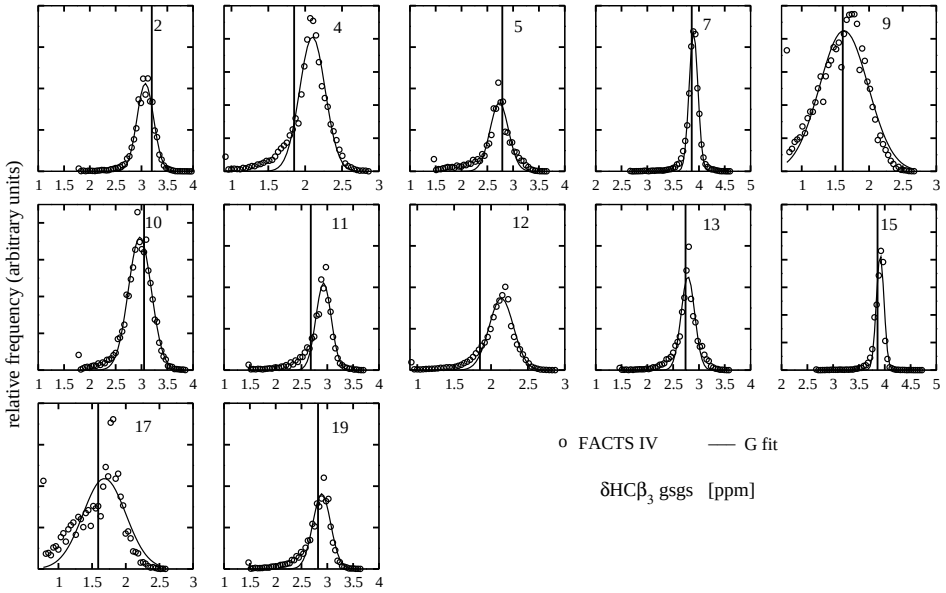


Figure 9.68: gsgs HCβ_3 CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS IV. Vertical lines represent the experimental shifts. See Tab. 9.16, Tab. 9.17 and Fig. 9.69 for quantitative analysis.

res.	exp. HCβ	FACTS sim. HCβ	I σHCβ	FACTS sim. HCβ	II σHCβ	FACTS sim. HCβ	III σHCβ	FACTS sim. HCβ	IV σHCβ
2	**3.20**	3.08	0.14	3.08	0.16	3.06	0.16	3.07	0.16
2	**3.20**	3.31	0.15	3.29	0.16	3.31	0.17	3.32	0.16
3	**1.56**	1.94	0.04	1.90	0.21	1.90	0.18	1.90	0.20
4	**1.85**	2.08	0.08	2.07	0.11	2.09	0.16	2.09	0.18
4	**1.94**	2.08	0.05	2.07	0.13	2.08	0.18	2.09	0.20
5	**2.79**	2.68	0.06	2.73	0.13	2.71	0.16	2.73	0.18
5	**2.79**	2.72	0.08	2.72	0.14	2.73	0.15	2.73	0.18
7	**3.86**	3.82	0.10	3.90	0.07	3.89	0.09	3.90	0.08
7	**3.93**	3.84	0.07	3.86	0.08	3.86	0.09	3.87	0.10
8	**4.20**	4.07	0.37	4.26	0.23	4.23	0.17	4.25	0.16
9	**1.61**	1.48	0.35	1.64	0.45	1.61	0.35	1.62	0.37
9	**1.61**	1.56	0.21	1.69	0.27	1.63	0.24	1.65	0.24
10	**3.05**	2.98	0.14	2.99	0.20	2.99	0.19	2.97	0.23
10	**3.13**	3.29	0.18	3.31	0.20	3.29	0.20	3.29	0.21
11	**2.68**	2.89	0.10	2.90	0.12	2.93	0.14	2.93	0.15
11	**2.75**	3.00	0.09	2.96	0.13	3.01	0.16	3.01	0.19
12	**1.85**	2.10	0.06	2.10	0.17	2.12	0.13	2.14	0.17
12	**1.96**	2.11	0.05	2.16	0.17	2.14	0.15	2.17	0.18
13	**2.74**	2.76	0.06	2.79	0.09	2.79	0.12	2.78	0.12
13	**2.82**	2.72	0.06	2.82	0.15	2.78	0.13	2.78	0.14
15	**3.86**	3.82	0.05	3.91	0.06	3.91	0.07	3.91	0.07
16	**4.20**	4.25	0.03	4.28	0.09	4.27	0.07	4.28	0.09
17	**1.59**	1.55	0.33	1.59	0.46	1.73	0.23	1.67	0.33
17	**1.59**	1.65	0.16	1.69	0.24	1.72	0.15	1.70	0.20
18	**1.65**	1.85	0.05	1.93	0.15	1.89	0.15	1.92	0.19
19	**2.82**	2.87	0.09	2.86	0.17	2.88	0.16	2.90	0.17
19	**3.10**	2.98	0.06	2.89	0.18	2.94	0.17	2.92	0.20
20	**4.27**	4.30	0.02	4.30	0.09	4.29	0.06	4.31	0.09

Table 9.16: Comparison between experimental (bold) and simulated δHCβ CS of gsgs with FACTS (at 300 K).

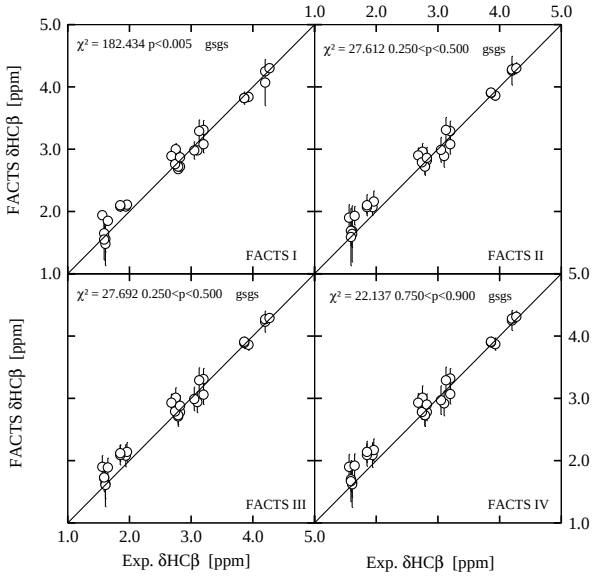


Figure 9.69: HCβ CS of FACTS simulations with gsgs in comparison with experimental values (300 K).

par.	DF	χ^2	p
FACTS I	28	182.434	p¡0.005
FACTS II	28	27.6118	0.250¡p¡0.500
FACTS III	28	27.6917	0.250¡p¡0.500
FACTS IV	28	22.137	0.750¡p¡0.900

Table 9.17: Statistical analysis of gsgs δ HCβ shifts (at 300 K) for experimental and calculated values.

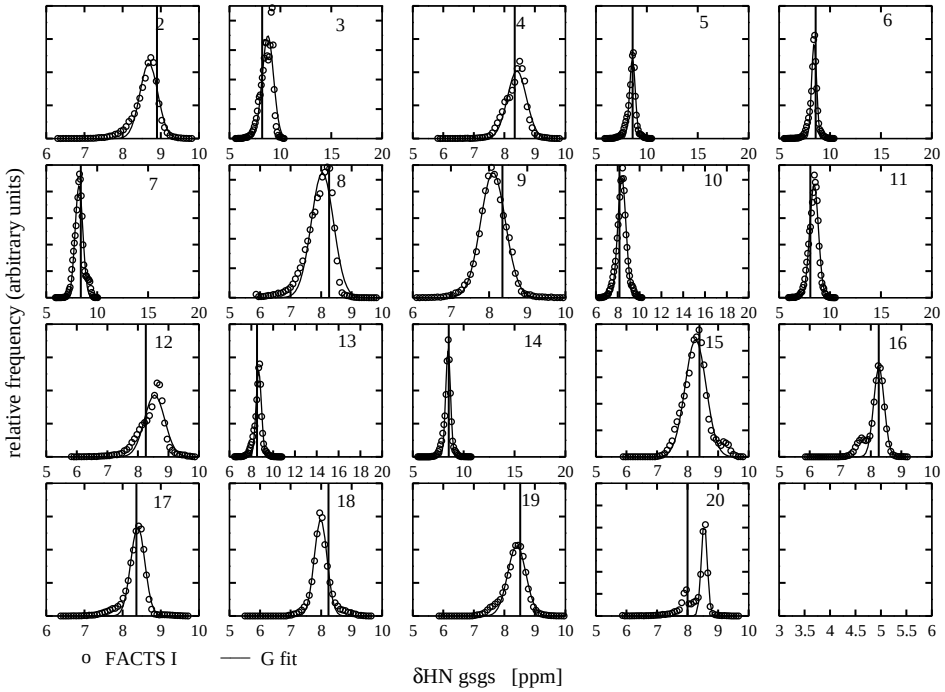


Figure 9.70: gsgs HN CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS I. Vertical lines represent the experimental shifts. See Tab. 9.18, Tab. 9.19 and Fig. 9.74 for quantitative analysis.

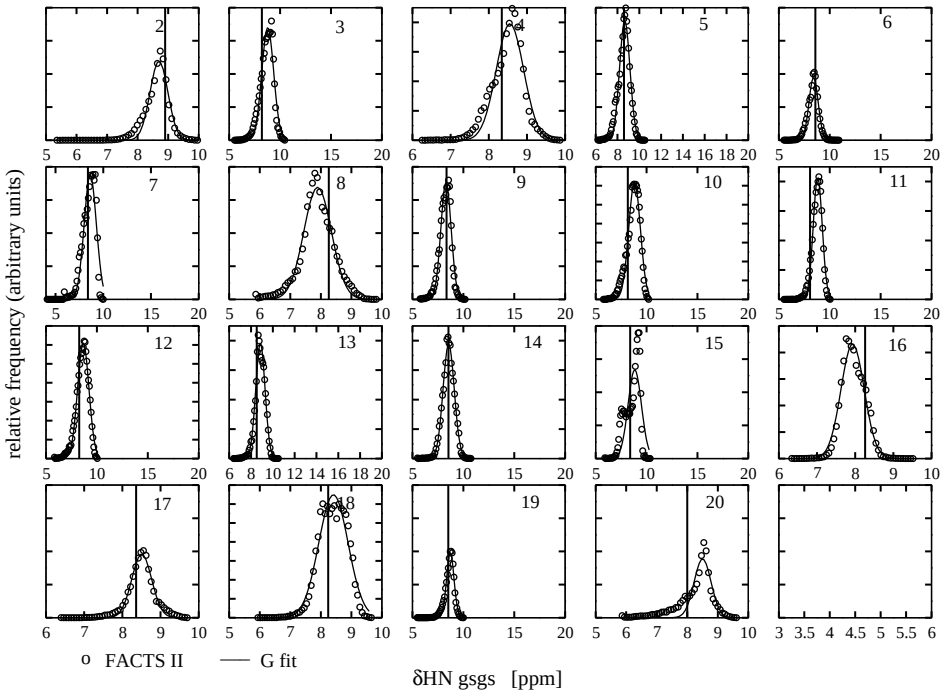


Figure 9.71: gsgs HN CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS II. Vertical lines represent the experimental shifts. See Tab. 9.18, Tab. 9.19 and Fig. 9.74 for quantitative analysis.

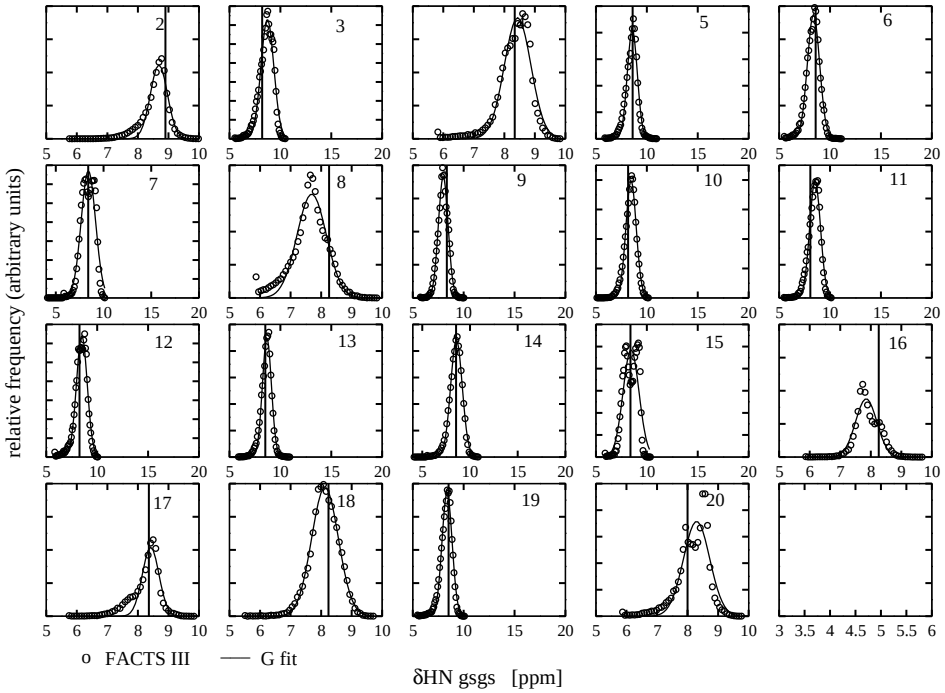


Figure 9.72: gsgs HN CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS III. Vertical lines represent the experimental shifts. See Tab. 9.18, Tab. 9.19 and Fig. 9.74 for quantitative analysis.

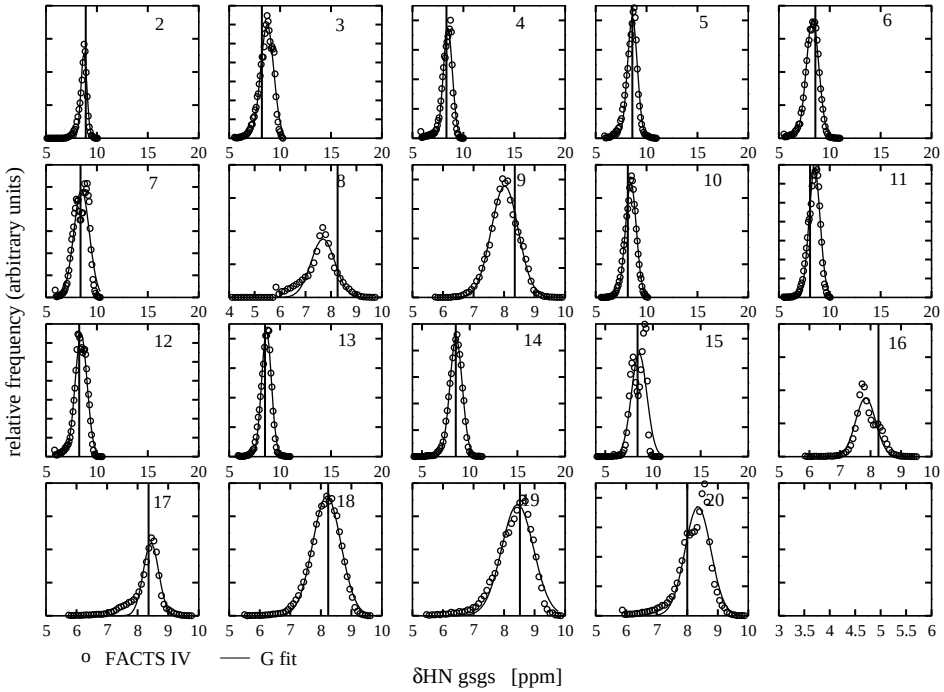


Figure 9.73: gsgs HN CS calculated via SHIFTX program from the FACTS simulations (6 μs) at 300 K with FACTS IV. Vertical lines represent the experimental shifts. See Tab. 9.18, Tab. 9.19 and Fig. 9.74 for quantitative analysis.

res.	exp. HN	FACTS sim. HN	I σHN	FACTS sim. HN	II σHN	FACTS sim. HN	III σHN	FACTS sim. HN	IV σHN
2	**8.90**	8.69	0.23	8.70	0.29	8.69	0.28	8.72	0.29
3	**8.21**	8.77	0.56	8.78	0.59	8.72	0.64	8.73	0.64
4	**8.34**	8.41	0.31	8.55	0.36	8.44	0.44	8.48	0.44
5	**8.60**	8.59	0.28	8.67	0.48	8.57	0.49	8.61	0.48
6	**8.58**	8.43	0.25	8.41	0.45	8.37	0.60	8.31	0.62
7	**8.39**	8.21	0.38	8.64	0.64	8.41	0.76	8.49	0.77
8	**8.26**	8.04	0.36	7.89	0.45	7.69	0.48	7.68	0.47
9	**8.35**	8.10	0.33	8.30	0.51	8.00	0.48	8.02	0.44
10	**8.16**	8.38	0.38	8.82	0.57	8.45	0.51	8.46	0.53
11	**8.06**	8.43	0.43	8.81	0.44	8.53	0.53	8.51	0.55
12	**8.26**	8.55	0.33	8.66	0.56	8.49	0.53	8.45	0.59
13	**8.52**	8.68	0.25	8.86	0.43	8.67	0.44	8.66	0.48
14	**8.54**	8.47	0.24	8.56	0.55	8.58	0.61	8.58	0.65
15	**8.39**	8.26	0.34	8.84	0.66	8.45	0.79	8.55	0.81
16	**8.26**	8.27	0.17	7.93	0.30	7.85	0.35	7.84	0.32
17	**8.36**	8.40	0.18	8.51	0.26	8.42	0.27	8.44	0.25
18	**8.24**	7.98	0.21	8.41	0.48	8.13	0.44	8.20	0.46
19	**8.52**	8.41	0.30	8.72	0.36	8.37	0.50	8.44	0.52
20	**8.00**	8.55	0.10	8.50	0.28	8.30	0.43	8.35	0.42

Table 9.18: Comparison between experimental (bold) and simulated δHN CS of gsgs with FACTS (at 300 K).

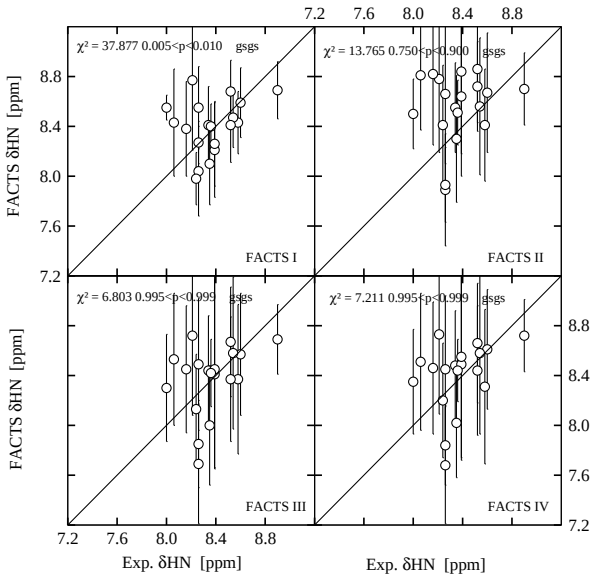


Figure 9.74: HN CS of FACTS simulations with gsgs in comparison with experimental values (300 K).

par.	DF	χ^2	p
FACTS I	19	37.8765	0.005¡p¡0.010
FACTS II	19	13.765	0.750¡p¡0.900
FACTS III	19	6.80317	0.995¡p¡0.999
FACTS IV	19	7.21109	0.995¡p¡0.999

Table 9.19: Statistical analysis of gsgs δ HN shifts (at 300 K) for experimental and calculated values.

exp. NOEs/FACTS	I	II	III	IV
very weak 4.0-6.0				
HC$\beta\beta'$W2-HCϵ_3Y11	5.33	5.18	4.86	4.77
HCδY11-HCγ_2I18	5.89	4.10	4.95	4.73
HNT20-H_3CγI18	5.93	6.10	5.77	5.77
violations:	**0**	**1**	**0**	**0**
weak 3.5-5.5				
HCαT1-HCϵY11	8.32	7.09	6.70	6.14
H_3CϵW2-HCαY11	5.07	3.95	4.58	4.55
H_3CϵW2-HCβ'N13	7.59	7.43	4.84	5.02
H_1CδW10-HCβ'Q12	6.56	5.89	6.21	6.29
HCδY11-HCαN13	5.75	5.12	5.78	5.87
HCδY11-HCγ'I18	5.42	4.05	4.52	4.42
HCϵY11-HC$\delta\delta'$K9	3.53	3.57	3.91	3.81
violations:	**4**	**3**	**3**	**3**
medium 2.5-4.5				
HCδY11-H_3CγI18	5.99	4.35	4.65	4.41
HCϵY11-HCαN13	5.39	4.67	5.28	5.30
HCϵY11-HCβ'N13	6.74	5.89	5.62	5.63
HCϵY11-H_3CγI18	8.07	5.52	5.95	5.79
HNT16-HCα'G14	4.96	4.03	4.15	4.08
HC$\beta\beta'$K17-HCϵY19	3.78	3.54	3.98	3.87
violations:	**5**	**3**	**4**	**3**
medium strong 2-4				
HCαW2-HCαY11	5.91	3.97	4.27	3.91
HCαW10-HCαY19	4.99	3.00	3.99	3.74
HCαQ12-HCαK17	3.40	2.49	2.70	2.67
violations:	**2**	**0**	**1**	**0**
strong 1.5-3.5				
HCαQ4-HCαK9	3.46	2.79	3.06	2.98
violations:	**0**	**0**	**0**	**0**

Table 9.20: Violations of the medium- and long-range NOE connectivities of the gsgs peptide, related to FACTS with different parametrisations. Experimental data are related to 1 mM of gsgs peptide in aqueous solution, pH 3.4, at 10 C°. Simulations are 3 μs long. See Table 1 of [50] for details.

Small proteins

PDB id.	struct.	n. res.	RMSD
1vii	NMR	36	$4 \div 33$
2cyu	NMR	39	$3 \div 14$
			$19 \div 38$
1crn	X-ray	46	$1 \div 35$
1enh	X-ray	54	$7 \div 54$
1igd	X-ray	61	$5 \div 61$
2ci2	X-ray	65	$5 \div 35$
			$46 \div 65$
2a3d	NMR	73	$1 \div 73$
1ubq	X-ray	76	$1 \div 17$
			$22 \div 50$
			$55 \div 72$
1pht	X-ray	83	$1 \div 9$
			$25 \div 39$
			$43 \div 80$

Table 9.21: The FACTS small protein test-case. The rightmost column shows the residues involved in the C_α-RMSD calculations.

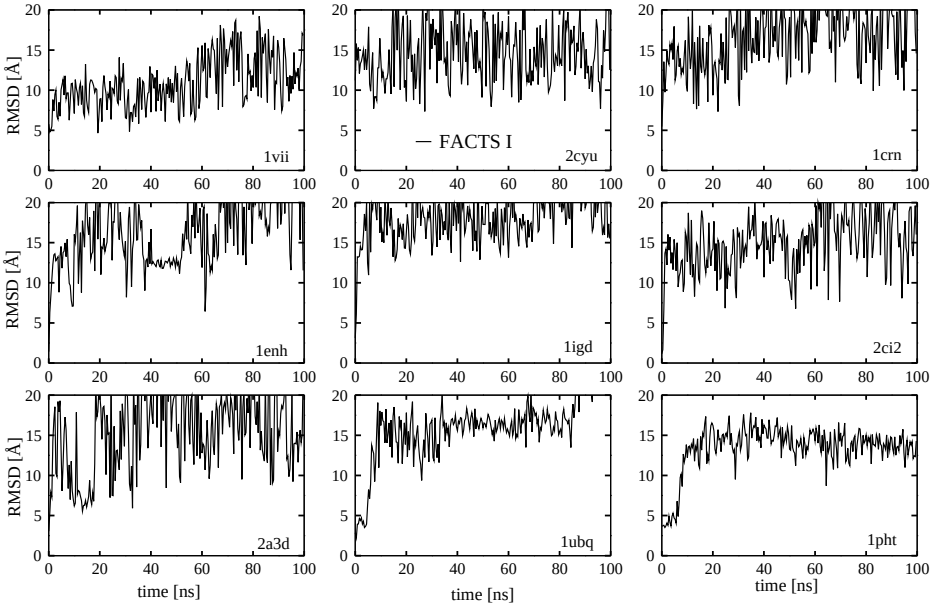


Figure 9.75: RMSD timeseries of small protein testcase with FACTS I.

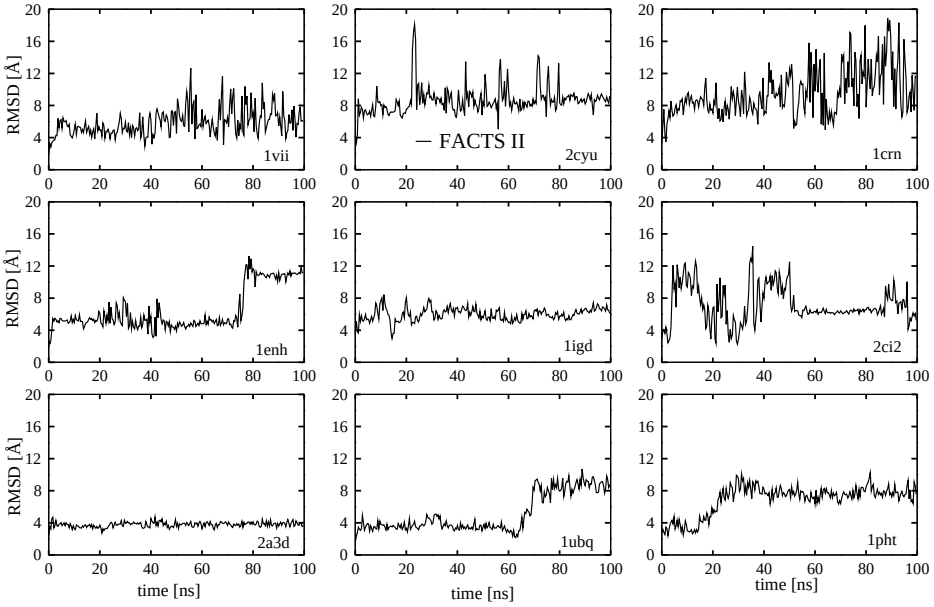


Figure 9.76: RMSD timeseries of small protein testcase with FACTS II.

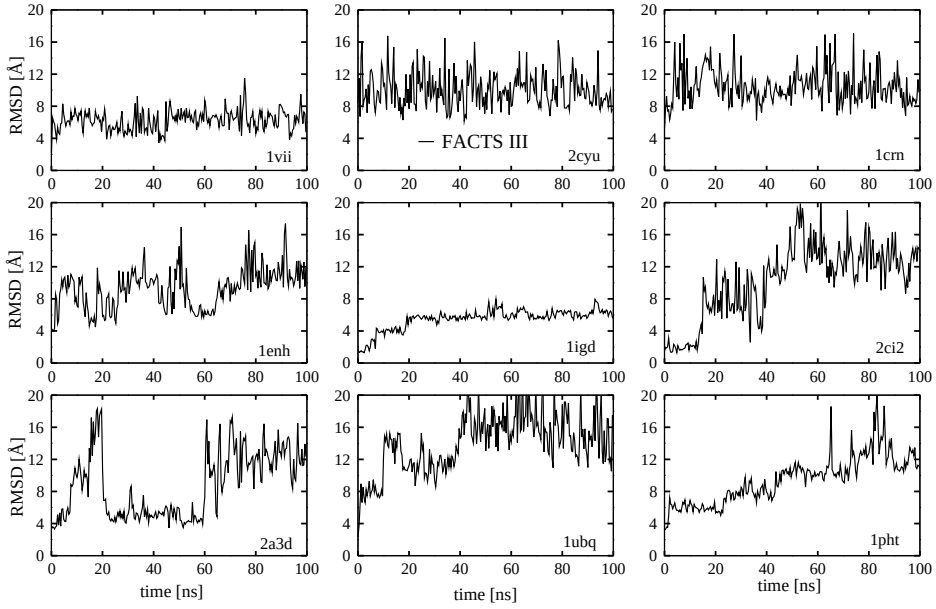


Figure 9.77: RMSD timeseries of small protein testcase with FACTS III.

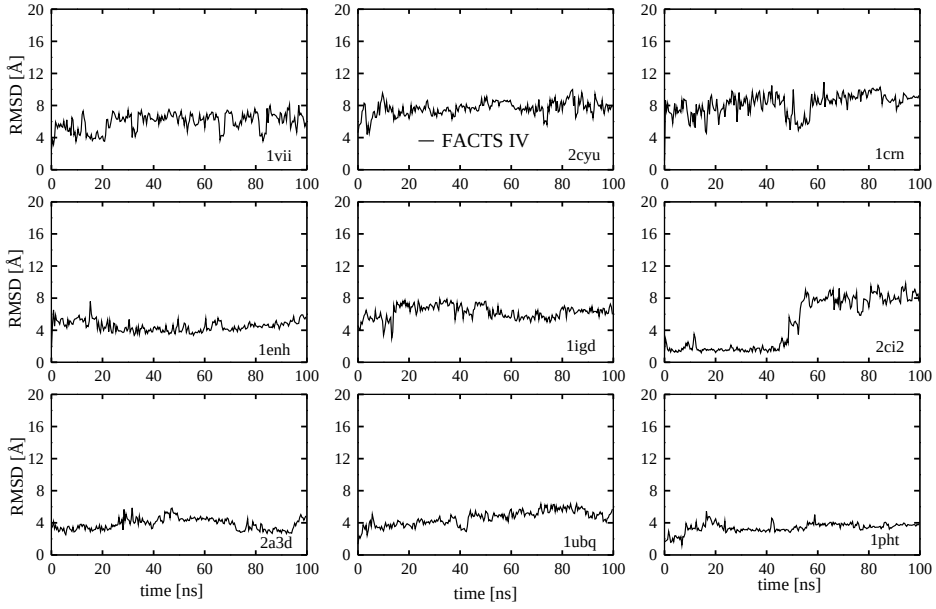


Figure 9.78: RMSD timeseries of small protein testcase with FACTS IV.

p./ns	10	20	30	40	50	60	70	80	90	100
1vii	11.3	9.0	8.7	8.5	11.0	17.3	14.9	11.4	17.0	16.9
2cyu	16.4	6.6	17.4	16.1	24.9	16.0	11.9	15.0	11.9	15.5
1crn	11.4	13.4	16.2	18.0	18.2	20.3	22.4	18.4	21.2	21.1
1enh	16.5	20.1	16.8	15.9	12.5	13.8	14.5	20.0	21.2	18.4
1igd	19.8	17.3	14.7	18.9	16.9	15.1	22.6	26.7	19.8	22.2
2ci2	15.1	16.1	18.1	12.9	13.2	16.5	16.6	13.9	18.0	18.4
2a3d	6.3	15.3	20.8	18.9	14.2	16.6	17.0	28.3	14.2	24.7
1ubq	18.8	18.0	14.2	16.8	16.0	16.8	17.3	17.2	19.1	23.6
1pht	13.3	15.5	14.0	14.3	10.1	14.4	15.7	12.4	14.0	13.3

Table 9.22: Small protein test-case RMSD timeseries table (FACTS I) for at 300 K.

p./ns	10	20	30	40	50	60	70	80	90	100
1vii	5.2	4.8	4.0	8.1	6.5	6.6	6.5	5.3	5.9	10.7
2cyu	7.8	7.4	7.2	7.6	8.5	9.4	8.1	7.6	8.7	8.8
1crn	8.6	6.7	9.6	12.0	8.8	10.7	9.0	14.0	10.0	7.1
1enh	5.2	6.6	5.9	5.6	4.2	5.2	4.9	12.5	10.9	11.4
1igd	7.9	8.0	7.0	6.7	4.9	5.2	5.8	5.4	6.2	7.5
2ci2	11.4	6.5	3.0	8.4	9.4	6.2	6.7	6.4	8.2	4.7
2a3d	3.4	3.5	3.9	3.8	3.7	3.8	4.0	4.1	3.6	3.1
1ubq	4.0	3.3	3.8	3.0	3.8	2.8	7.8	9.2	8.4	10.1
1pht	2.8	5.9	8.2	8.1	7.8	6.5	6.5	8.4	8.0	7.2

Table 9.23: Small protein test-case RMSD timeseries table (FACTS II) for at 300 K.

p./ns	10	20	30	40	50	60	70	80	90	100
1vii	6.4	5.8	5.4	6.7	6.8	8.3	7.9	4.7	8.1	5.3
2cyu	16.6	7.0	9.1	10.3	9.7	9.3	8.3	13.0	10.1	10.7
1crn	9.8	10.5	14.5	10.3	10.6	10.6	8.8	10.2	12.7	10.4
1enh	5.8	6.1	11.4	9.5	12.4	7.1	9.6	10.8	15.3	10.2
1igd	3.8	5.0	5.7	5.7	5.6	6.4	5.9	6.1	5.8	6.0
2ci2	2.5	7.0	6.8	11.2	13.3	14.1	11.5	19.3	12.1	12.7
2a3d	10.5	10.0	5.2	4.9	5.0	6.7	15.2	13.2	12.2	13.3
1ubq	11.2	12.4	8.6	15.1	12.6	15.5	18.8	14.6	11.3	11.4
1pht	7.0	5.9	9.2	8.8	9.8	10.4	9.4	13.1	10.5	10.5

Table 9.24: Small protein test-case RMSD timeseries table (FACTS III) for at 300 K.

p./ns	10	20	30	40	50	60	70	80	90	100
1vii	4.3	3.5	7.2	6.7	6.9	5.8	7.1	6.9	6.0	5.5
2cyu	9.2	7.8	6.4	7.8	8.7	8.4	7.2	7.7	8.2	6.8
1crn	5.1	7.4	7.7	8.7	5.8	8.9	9.8	9.4	9.0	9.2
1enh	5.5	4.4	3.8	4.7	3.8	4.0	4.1	4.7	5.1	6.2
1igd	3.8	6.9	6.8	7.5	7.6	5.5	6.3	6.7	6.1	6.8
2ci2	1.8	1.4	2.0	1.5	4.8	7.7	8.6	7.6	8.9	7.7
2a3d	3.3	3.3	4.1	3.8	4.5	4.2	3.8	3.9	3.0	5.7
1ubq	3.2	3.9	4.3	3.6	4.5	5.6	5.4	5.8	5.2	5.4
1pht	3.4	4.5	3.2	3.2	3.3	3.9	3.7	3.6	3.6	3.8

Table 9.25: Small protein test-case RMSD timeseries table (FACTS IV) for at 300 K.

Folding of the protein G

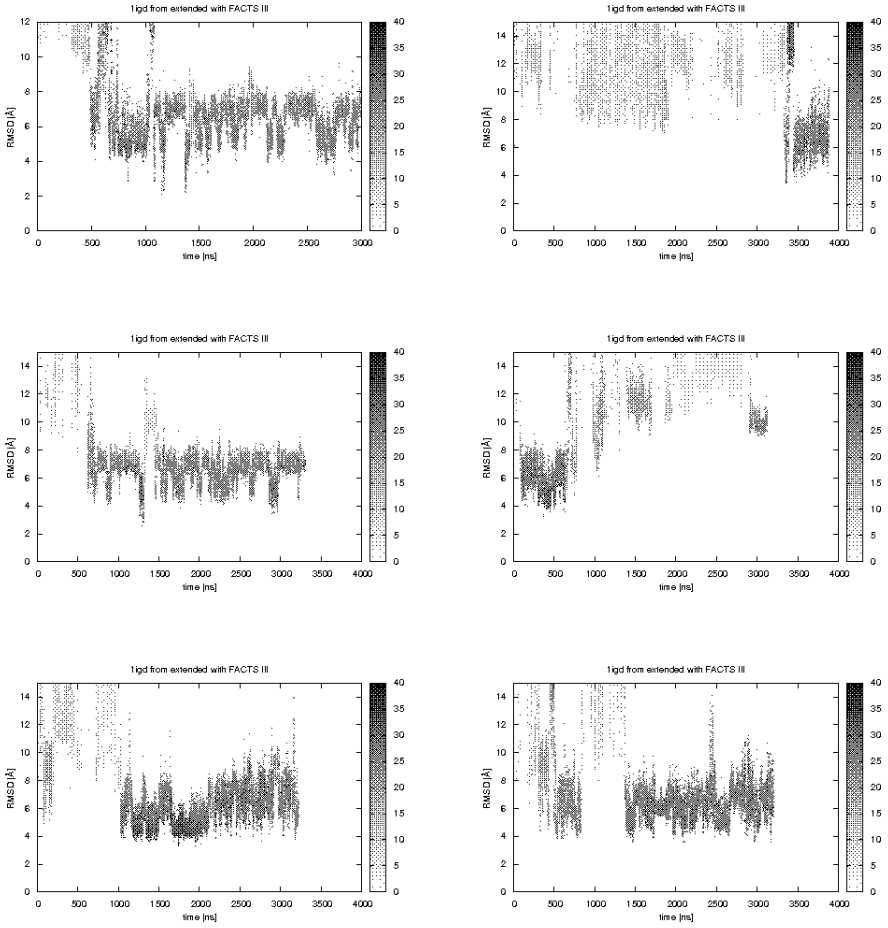


Figure 9.79: Time series of RMSD and contacts (black and white scale) extracted from six MD simulations (A, B, C, D, E, F, see next figure for an image of the most populated clusters) of the protein G at 300 K with the best parameter set (FACTS III) from extended.

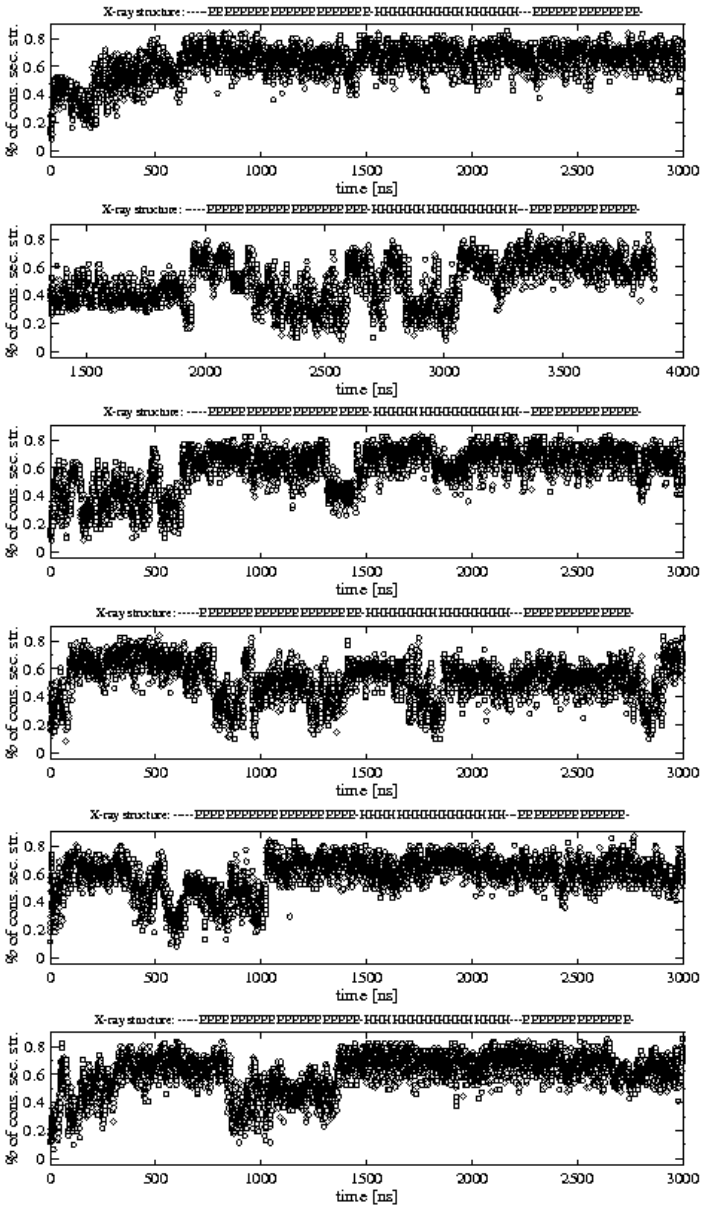
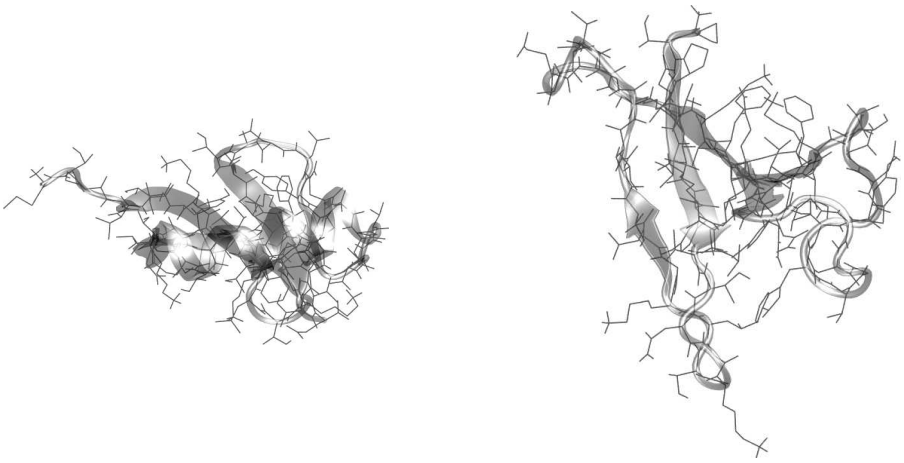
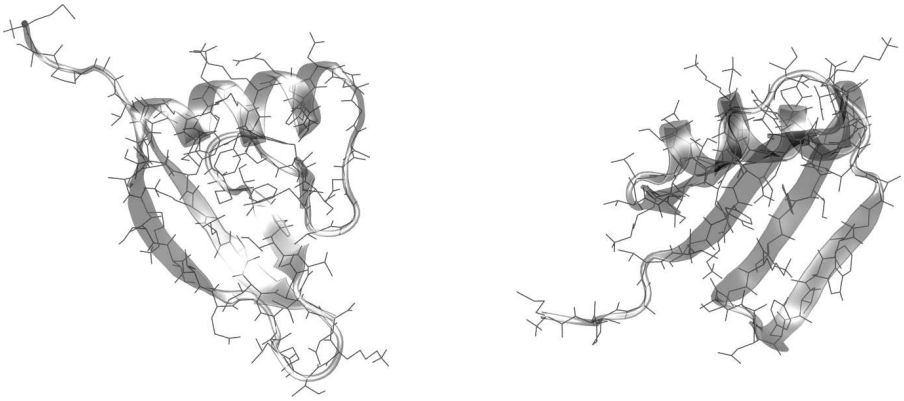


Figure 9.80: Time series of percent of conserved secondary structure along the simulations.



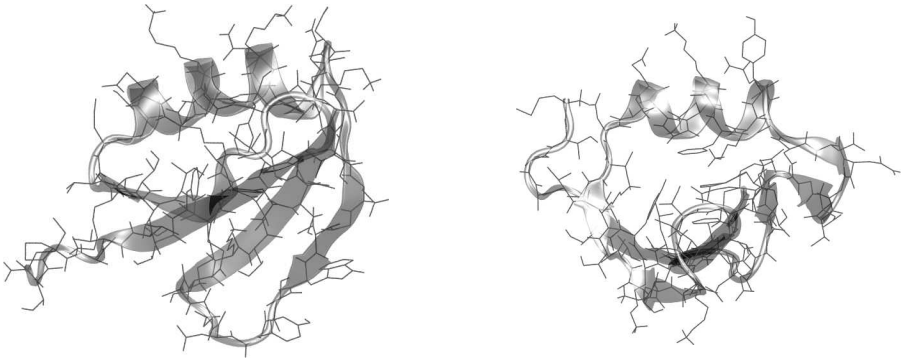
FACTS III (A) (pop. 0.77 %)

FACTS III (B) (pop. 8.55 %)



FACTS III (C) (pop. 7.09 %)

FACTS III (D) (pop. 1.16 %)



FACTS III (E) (pop. 3.58 %)

FACTS III (F) (pop. 2.40 %)

Figure 9.81: Central conformations of the first cluster of the four 1igd run (at 300 K) with FACTS III. The RMSD-clustering was performed with Wordom with a cutoff of 2.5 . Simulations are 3 μs long.

9.2 Appendix 2

9.2.1 FIGURES (from chapter 6)

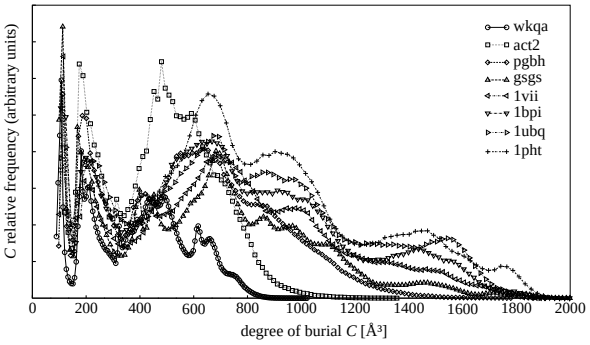


Figure 9.82: Examples of C distributions calculated for tyrosine hydroxylase 22-34 (wkqa), a three-stranded β sheets (gsgs), the C-terminal β hairpin of 1igd protein (pgbh), the helix Ace-$(AAAQAA)_3$-amide (act2), and four medium-size proteins (1vii, 1bpi, 1ubq, 1pht). Remarkably, between $C \in [1100, 1600]$ Å^3 a relative maximum is found for those structures which have a small hydrophobic core (e.g. 1vii, 1pht).

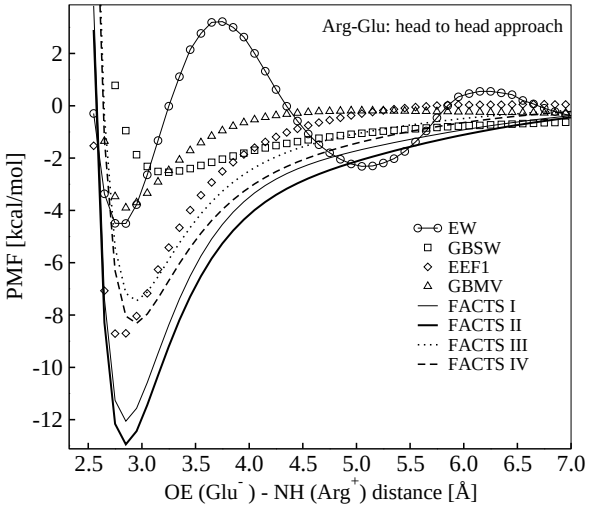


Figure 9.83: Explicit water PMF for arginine-glutamine head-to-head approach as published in [30] compared with different FACTS setups and other solvation models. FACTS III and IV, related to internal dielectric $\epsilon = 2$ give profiles which are closer to EW. Remarkably, FACTS II and IV profiles are close to EEF1 implicit solvent model as well. See Supplementary Material for a more detailed study of Lazaridis PMF.

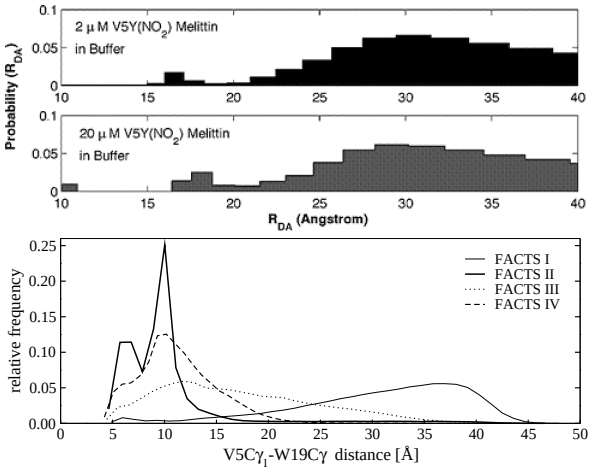


Figure 9.84: (Top) Donor-acceptor distance distributions from FET measurements of the V5Y(NO_2) melittin mutant. 2 μM peptide in 20 mM sodium phosphate buffer, pH 7.4 (black) and 20 μM peptide in 20 mM sodium phosphate buffer, pH 7.4 (grey) from Ref. [38]. (Bottom) Distributions of the Cγ_1Y5-CγW19 distance extracted from MD simulations with FACTS. Parameter sets I and III lead to dominant broad, single mode distributions, which are consistent with the random-coil configuration indicated by CD spectroscopy.

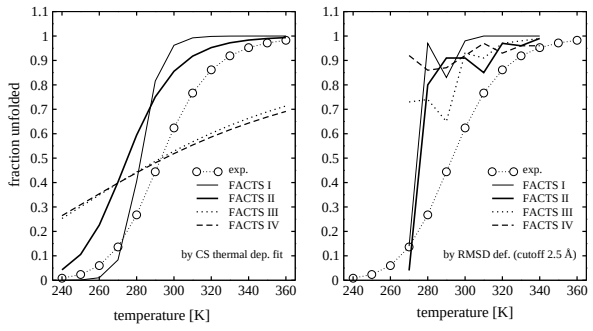
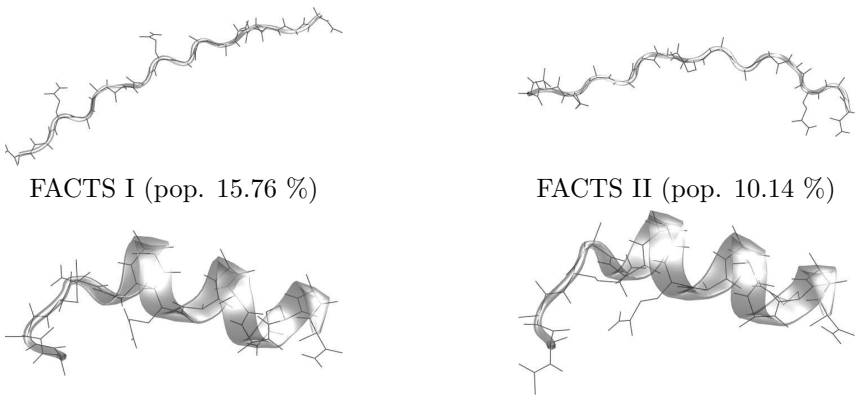


Figure 9.85: Comparison between Honda's trends of unfolded percentage of pgbh in water as a function of the temperature. Experimental data is obtained by means of thermal dependence of HCα shifts measurements (see Supp. Mat and Ref. [43] for details). FACTS trends are recovered by HCα shifts calculation (left) and assuming that a conformation is in an unfolded state if its RMSD with respect to NMR 1pgb exceeds 2.5 Å(right). From 4 μs MD simulations at 270, 280, 290...350 K.



FACTS I (pop. 15.76 %) FACTS II (pop. 10.14 %)

FACTS III (pop. 8.69 %) FACTS IV (pop. 7.50 %)

Figure 9.86: Central conformations of peptide act2 (at 274 K). The RMSD-clustering was performed with Wordom with a cutoff of 2.5 . Simulation time: 4 μs.

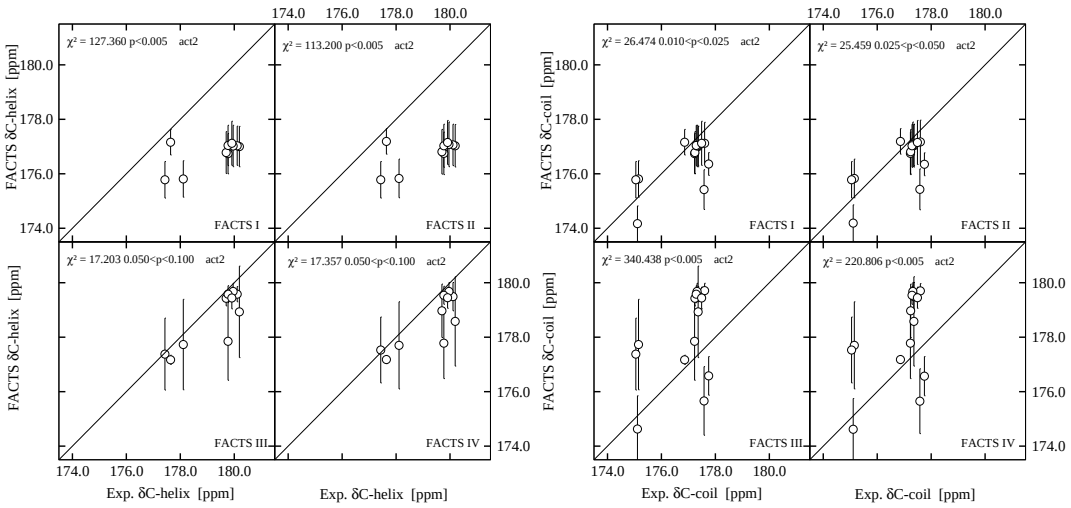


Figure 9.87: (Left) Comparison between act2 experimental carbonyl-carbon CS related to helical conformations of the peptide. (Right) The same comparison with coil conformations. FACTS data comes from the analysis of 4 μs MD simulations at 274. The best agreement with experimental data is obtained with parameter sets III and IV, which agree with experimental helix shifts at 10°C. FACTS I and II setups are closer to coil conformation, related to 90°C.

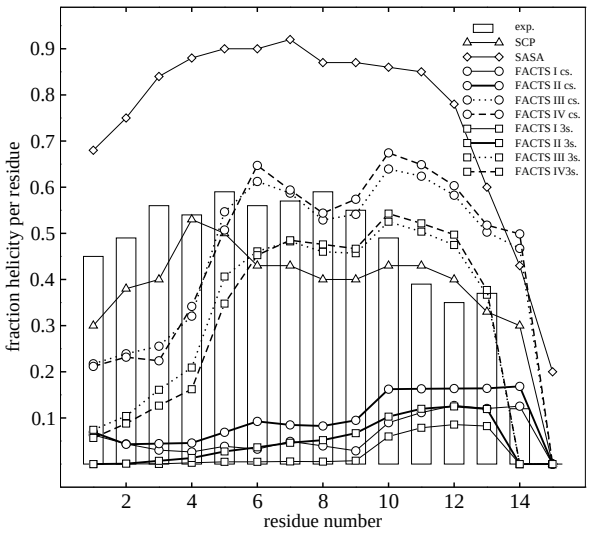


Figure 9.88: Comparison between Acetyl-$(AAQAA)_3$-NH_2 helicity content per residue obtained by Stellwagen and coworkers from two-state analysis of the thermal dependence of carbonyl-carbon CS measurements in pure water at 1°C (bars), Hassan and coworkers SCP implicit solvent model [48] (triangles), Caflisch and coworkers SASA implicit solvent model [67] (diamonds) and the one related to FACTS with the four different parametrisations via CS measurements (circles) and three-segments method (squares). Notably, FACTS showed a clear better behaviour with $\epsilon = 2$, closer to SCP model. Simulations with different values of γ yield approximately the same results. These outcomes confirm the δC trend seen in Fig. 9.87.

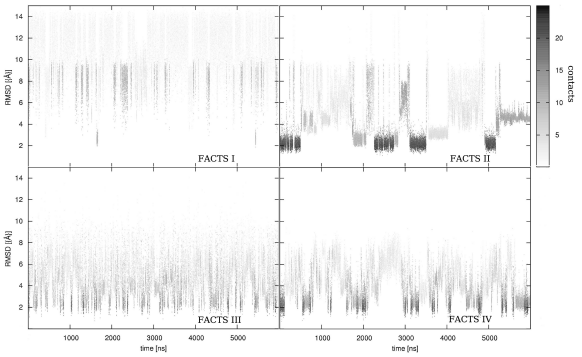
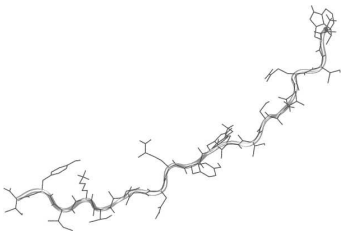
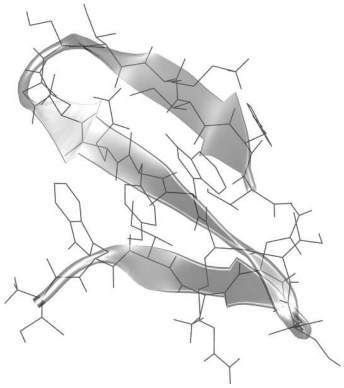


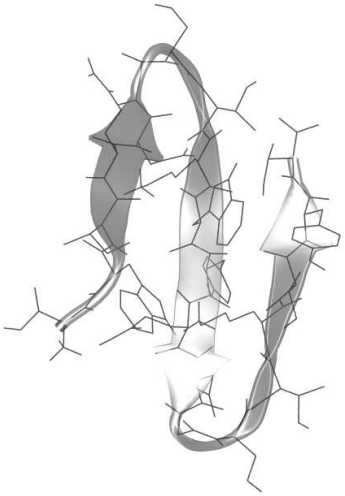
Figure 9.89: Time series of RMSD and contacts (black and white scale) of the gsgs peptide at 300 K with FACTS I, II, III IV. In this picture, a folding event occurs when a low, dark region is inserted between two high, light ones, or viceversa.



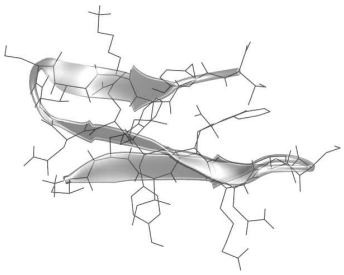
FACTS I (pop. 2.13 %)



FACTS II (pop. 19.52 %)



FACTS III (pop. 10.47 %)



FACTS IV (pop. 30.83 %)

Figure 9.90: Central conformations of peptide gsgs (at 300 K). The RMSD-clustering was performed with Wordom with a cutoff of 2.5 . Simulation time: 6 μs.

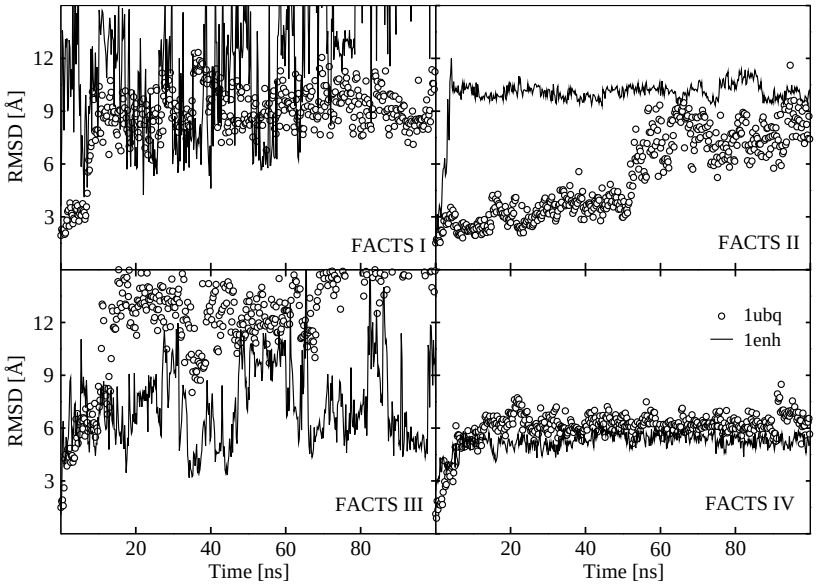


Figure 9.91: Time series of RMSD of ubiquitine (1ubq, line) and of the engrailed homeodomain of drosophila (circles) with different FACTS setups at 300 K. There comes out a tendency to instability, more pronounced for FACTS I and II, with internal dielectric $\epsilon = 1$. Simulation time: 100 ns.

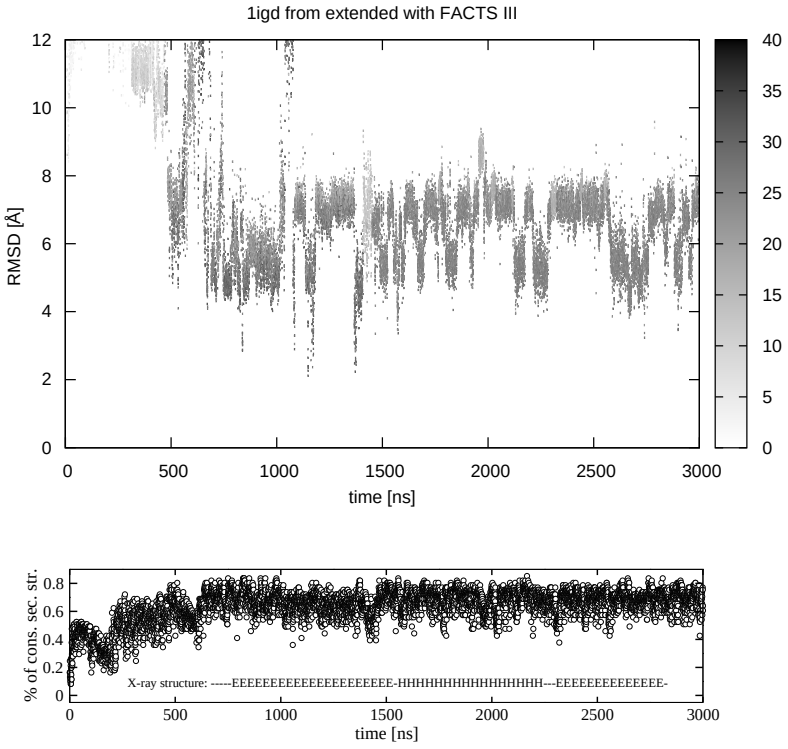


Figure 9.92: (Up) Timeseries of RMSD and contacts extracted from MD simulations of protein G at 300 K with one of the best parameter sets (FACTS III). (Bottom) Timeseries of percentage of secondary structure content with respect to crystal conformer 1igd. Simulation time: 3 μs.

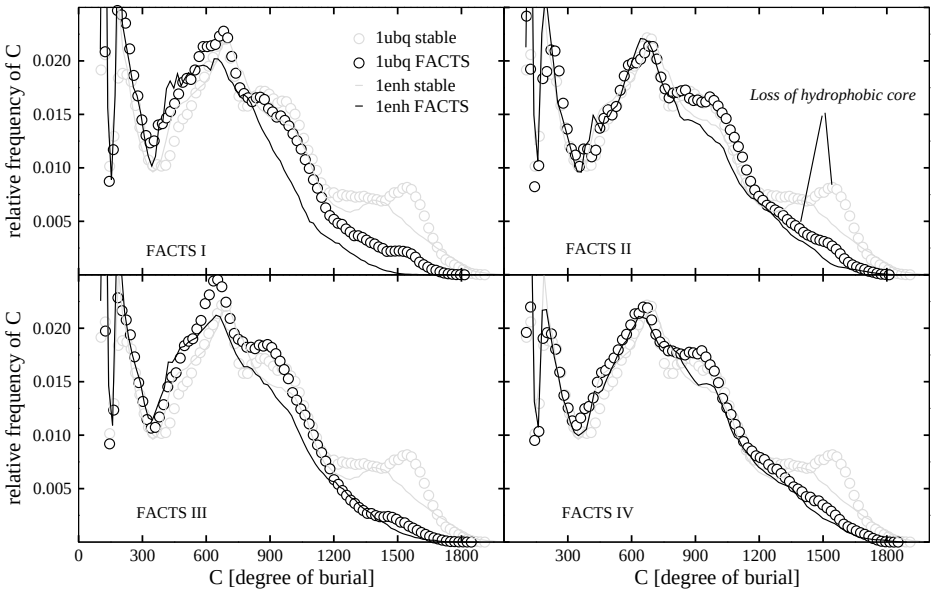


Figure 9.93: Loss of the hydrophobic core of 1ubq (+) and 1enh (∘) in relation with the same FACTS simulations in Fig. 9.91, seen through the C distributions of their atoms. Grey symbols refer to 100 ns of harmonic constraints dynamics, while other colours correspond to the 100 ns simulations of Fig. 9.91. By comparing grey and black distributions, one can see the peaks disappearing (around $C = 1500$ for 1ubq and $C = 1300$ for 1enh), more prominently for FACTS I and II. This means that the small hydrophobic cores of 1ubq and 1enh are lost during all the FACTS simulations. The other regions of the C distributions are mostly preserved.

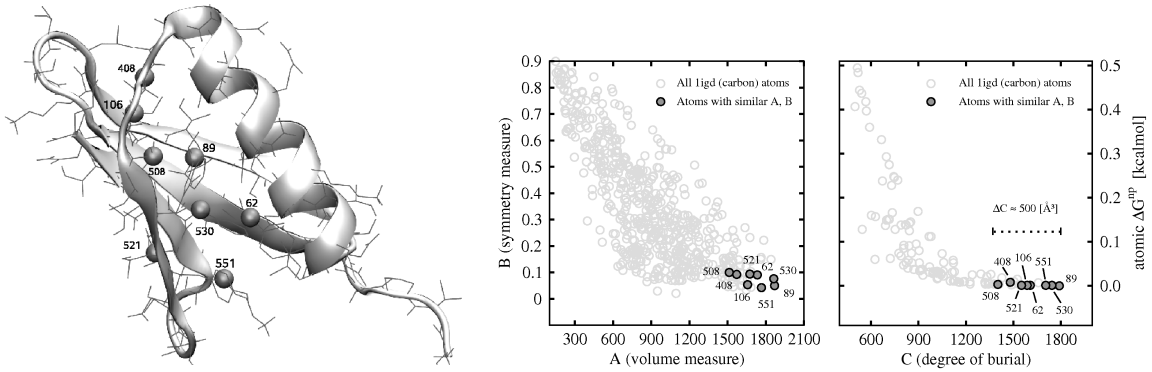


Figure 9.94: (Left) 1igd protein carbon atoms (in green) such that the volume measure $A_i \in [1550, 2000]$ (equal to $\simeq 17\%$ of the total A range) and the symmetry measure $B_i \in [0, 0.1]$ (equal to 10% of the total B range) (central plot). (Right plot) FACTS nonpolar (carbon atoms) contribution to solvation energy as a function of the degree of burial C. All the selected atoms give almost the same contribution, but their degree of burial covers over 500 Å^3 (equal to $\simeq$ 25 % of the total C range). The FACTS degree of burial C can thus be used to correct the nonpolar contribution. The more an atom is buried (like atoms 89, 530, and 551) the less its contribution to solvation energy should be. The more an atom is exposed (like atoms 408, 508 and 521) the more its contribution is close to the SASA case.

9.2.2 Supplementary Material (Chapter 6)

From Introduction

macromolecule	size	secondary structure	best internal ϵ
22-34 tyrosine hydroxylase [31, 32]	12	unstructured	low
Ace-$(AAQAA)_3$-amide [45]	15	α-helix	high
45-61 fragment of protein G [54]	16	β-hairpin	high/low
gsgs peptide [50]	20	3-str. β-sheets	high/low
melittin [35]	26	unstructured	low
small proteins [73](Tab. 9.28)	60-83	highly structured	high

Table 9.26: Details from Tab. 1.

FACTS PARAMETRISATIONS

$(\epsilon = 1, \gamma = 7.5) = \mathbf{I}$: strong (internal) electrostatic interactions, weak nonpolar interactions; expected to reduce the stability of globular structures, to keep unstructured peptides elongated and to slow down kinetics.

$(\epsilon = 1, \gamma = 15) = \mathbf{II}$: strong electrostatics, strong nonpolar interactions.

$(\epsilon = 2, \gamma = 7.5) = \mathbf{III}$: weak electrostatics, weak nonpolar interactions; setting $\epsilon = 2$ means to halve (in module) the electrostatics contribution to the solvation energy, since $\tau = \frac{78.5-2}{78.5 \cdot 2} \simeq 0.487$ while $\tau = \frac{78.5-1}{78.5 \cdot 1} \simeq 0.988$.

$(\epsilon = 2, \gamma = 15) = \mathbf{IV}$: weak electrostatics, strong nonpolar interactions.

FACTS SISI PARAMETRISATIONS

set id.	g_0	g_1	g_2	g_3
A	0	-0.6	0.01	1300
B	0	-0.6	0.02	1400
C	0	-0.7	0.01	1300
D	0	-0.7	0.02	1400

Table 9.27: Parametrisation sets indices for FACTS SISI correction.

From Results and Discussion

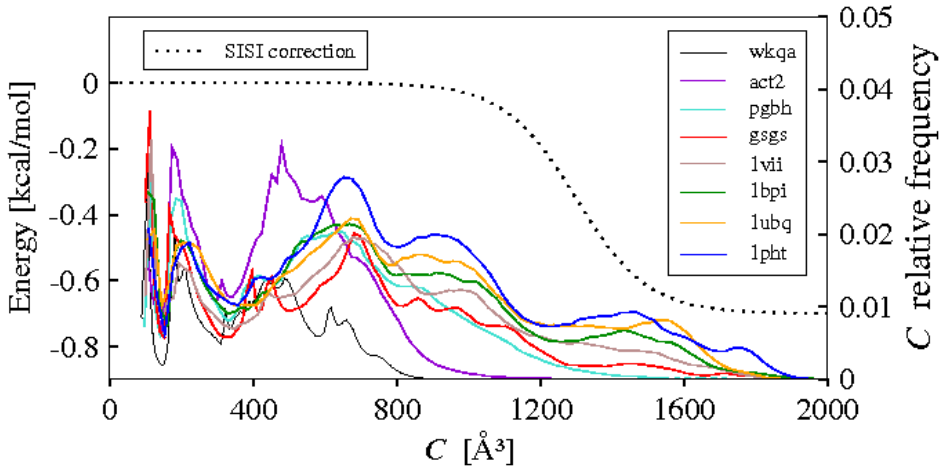


Figure 9.95: Examples of FACTS degree-of-burial-frequency distributions (C) of all atoms belonging to some structure within the FACTS SISI testcase (obtained with 100 ns Berendsen dynamics simulations at 300 K). While wkqa, act2 and pgbh distributions decrease with C monotonically, reaching 0 frequency around $C = 800 - -1400$, the proteins show a different behaviour, since many atoms are buried within the structures. In particular, 1ubq (orange) shows a relative maximum at about $C = 1600$. Remarkably, the gsgs peptide (red) and the villin headpiece 1vii (brown) can be placed between these two different trends (the gsgs shows a very smoothed relative maximum around $C = 1500$, while the villin headpiece smoothly decreases over $C = 1400$) and, therefore, they can be thought of as intermediate cases between extended, well-solvated structures (such as wkqa or act2) and globular proteins (such as 1ubq or 1pht). The black, dotted line is the FACTS SISI correction of the nonpolar contribution to solvation energy. It is important to point out that the correction affects mainly the structures whose C-distributions show a hydrophobic core (like 1ubq or 1pht) or, at least, a relative maximum beyond $C = 1200$.

	struct.	n. of res.	RMSD on which C_α
1vii	NMR	36	$4 \div 33$
2cyu	NMR	39	$3 \div 14$, $19 \div 38$
1enh	X-ray	54	$7 \div 54$
1bpi	X-ray	58	$2 \div 7$, $17 \div 36$, $47 \div 57$
1igd	X-ray	61	$5 \div 61$
2ci2	X-ray	65	$5 \div 35$, $46 \div 65$
2a3d	NMR	73	$1 \div 73$
1ubq	X-ray	76	$1 \div 17$, $22 \div 50$, $55 \div 72$
1pht	X-ray	83	$1 \div 9$, $25 \div 39$, $43 \div 80$

Table 9.28: The FACTS SISI test-case. The second column is about which kind of technique was used to recover the structure (nuclear magnetic resonance or crystallisation) and the third one contains the number of residues of each structure and the last one shows which C_α-atom has been selected to calculate the RMSD (See Fig. 9.95).

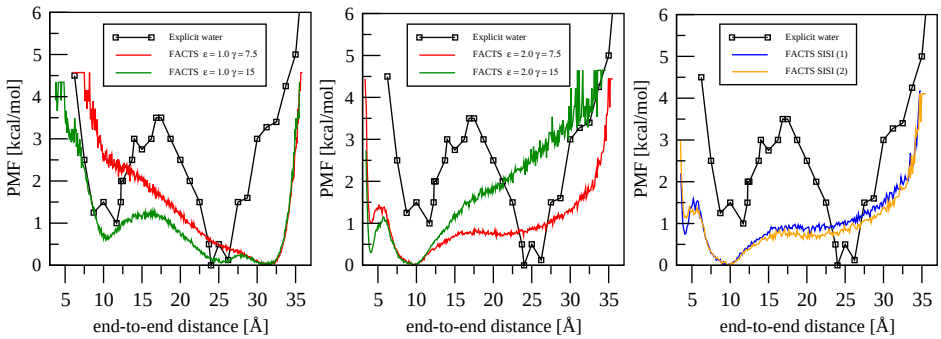


Figure 9.96: End-to-end distance PMFs of Stulz' peptide (wkqa) related to FACTS with different internal dielectric and surface tension (left and middle) and FACTS SISI (right, two runs of Tab. 6.3) in comparison with Stult'z explicit water data (black). Note that FACTS SISI profiles copy FACTS with $\epsilon = 2.0$, $\gamma = 7.5$, since the SISI correction does not affect this peptide too much. Although the explicit water minimum around 25 Å is not well shaped, the position of the first one around 10 Å has been achieved by FACTS and FACTS SISI.

Bibliography

[1] M. Feig and C. L. Brooks III, Curr. Opin. Struct. Biol. **14**, 217 (2004).

[2] D. Bashford and D. A. Case, Annu. Rev. Phys. Chem. **51**, 129 (2000).

[3] M. C. Lee, R. Yang, and Y. Duan, J. Mol. Model **12**, 101 (2005).

[4] J. Chen, W. Im, and C. L. Brooks III, J. Am. Chem. Soc. **128**, 3728 (2006).

[5] J. A. Wagoner and N. A. Baker, J. Comput. Chem. **25**, 1623 (2004).

[6] E. E. Meyer, K. J. Rosenberg, and J. Israelachvili, Proc. Natl. Acad. Sci. USA. **Early Edition**, 1 (2004).

[7] P. Labute, J. Comput. Chem. **29**, 1693 (2007).

[8] M. Zacharias, J. Phys. Chem. A **107**, 3000 (2003).

[9] J. Chen, C. L. Brooks III, and J. Khandogin, Curr. Opin. Struct. Biol. **18**, 140 (2008).

[10] U. Haberthuer and A. Caflisch, J. Comput. Chem. **29**, 701 (2008).

[11] O. A., B. D., and C. D. A., J. Comput. Chem. **23**, 1297 (2002).

[12] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, J. Am. Chem. Soc. **112**, 6127 (1990).

[13] B. R. Brooks *et al.*, J. Comput. Chem. **4**, 187 (1983).

[14] L. Smith, *Caos* (Oxford University Press, 2008).

[15] E. Lyman and D. M. Zuckerman, Biophys. J. **91**, 164 (2006).

[16] X. Daura, W. F. van Gunseteren, and A. E. Mark, Prot. Str. Fun. Bio. **34**, 269 (1999).

[17] L. J. Smith, X. Daura, and W. F. van Gunseteren, Prot. Str. Fun. Bio. **48**, 487 (2002).

[18] W. F. van Gunseteren, R. Buergi, and C. P. adn X. Daura, Angew. Chem. Int. Ed. **40**, 352 (2001).

[19] A. R. Dinner and M. Karplus, Angew. Chem. Int. Ed. **40**, 4615 (2001).

[20] W. F. van Gunseteren, R. Buergi, and C. P. adn X. Daura, Angew. Chem. Int. Ed. **40**, 4616 (2001).

[21] M. Seeber, M. Cecchini, F. Rao, G. Settanni, and A. Caflisch, Bioinformatics **23**, 2625 (2007).

[22] P. Ferrara and A. Caflisch, J. Mol. Biol. **306**, 837 (2001).

[23] A. Cavalli, U. Haberthür, E. Paci, and A. Caflisch, Protein Science **12**, 1801 (2003).

[24] S. Neal, A. M. Nip, H. Zhang, and D. S. Wishart, Journal of Biomolecular NMR **26**, 215 (2003).

[25] D. van der Spoel, Biochem. Cell Biol. **76**, 164 (1998).

[26] D. A. Case, C. Scheurer, and R. Brschweiler, J. Am. Chem. Soc. **122**, 10390 (2000).

[27] A. Bagno, F. D'Amico, and G. Saielli, ChemPhysChem **8**, 873 (2007).

[28] M. V. Berjanskii and D. S. Wishart, Journal of Biomolecular NMR **40**, 31 (2008).

[29] K. Seidel, M. Etzkorn, R. Schneider, C. Ader, and M. aldus, Solid State NMR (Accep. August 2008).

[30] A. Masunov and T. Lazaridis, J. Am. Chem. Soc. **125**, 1722 (2003).

[31] C. Stultz, A. D. Levin, and E. R. Edelman, J. Biol. Chem. **265**, 47653 (2002).

[32] C. Stultz, J. Phys. Chem. B **108**, 16525 (2004).

[33] J. R. Lakowicz, *Principles of Fluorescence Spectroscopy* (Plenum Press, New York, 1983).

[34] C. L. Brooks III, M. Karplus, and B. M. Pettitt, *Proteins: a theoretical perspective of dynamics, structure and thermodynamics, in Adv. Chem. Phys. LXXI* (John Wiley & Sons, New York, 1988).

[35] J. Lauterwein, L. R. Brown, and K. Wuethrich, Biochim. Biophys. Acta **622**, 219 (1980).

[36] A. Pastore, T. S. Harvey, C. Dempsey, and I. D. Campbell, Eur. Biophys. J. **16**, 363 (1989).

[37] M. Khajehpour, T. Troxler, V. Nanda, and J. Vanderkooi, Prot. Str. Fun. Bio. **55**, 275 (2004).

[38] R. Hartings, H. B. Gray, and J. R. Winkler, J. Phys. Chem. B **112**, 3203 (2008).

[39] C. A. F. Andersen, A. G. Palmer, S. Brunak, and B. Rost, Structure **10**, 174 (2002).

[40] A. Glaettli, I. Chandrasekhar, and W. F. van Gunsteren, Eur. Biophys. J. **35**, 255 (2005).

[41] J. C. Talbot, J. Dufourcq, J. de Bony, J. F. Faucon, and C. Lussan, FEBS Lett. **102**, 191 (1979).

[42] F. J. Blanco, G. Rivas, and L. Serrano, Nature Struct. Biol. **1**, 584 (1994).

[43] S. Honda, N. Kobayashi, and E. Munekata, J. Mol. Biol. **295**, 269 (2000).

[44] K. P. Murphy and S. J. Gill, J. Mol. Biol. **222**, 699 (1991).

[45] W. Sholongo, L. Dugad, and E. Stellwagen, J. Am. Chem. Soc. **116**, 8288 (1994).

[46] D. W. Marquardt, J. Soc. Ind. Appl. Math. **11**, 431 (1963).

[47] S. Sung and X. Wu, Prot. Str. Fun. Bio. **25**, 202 (1996).

[48] S. A. Hassan and E. L. Mehler, Int. J. Quantum Chem. **83**, 193 (2001).

[49] P. Ferrara and A. Caflisch, Prot. Str. Fun. Bio. **46**, 24 (2002).

[50] E. D. Alba, J. Santoro, M. Rico, and A. Jimenez, J. Am. Chem. Soc. **123**, 2970 (2001).

[51] E. D. Alba, M. A. Jimenez, M. Rico, and J. L. Nieto, Folding & Design **1**, 133 (1996).

[52] K. Wuetrich, M. Billeter, and W. Braun, J. Mol. Biol. **180**, 715 (1984).

[53] A. Prunotto, FACTS: A Systematic Study , Manuscript in preparation (2010).

[54] W. A. Maltese and J. D. Robishaw, J. Biol. Chem. **265**, 18071 (1990).

[55] D. Eisenberg and A. McLachlan, Nature **319**, 199 (1986).

[56] T. Ooi, M. Oobatake, G. Nmethy, and H. A. Scheraga, Proc. Natl. Acad. Sci. USA. **84**, 3086 (1987).

[57] F. Fraternali and W. F. Van Gunsteren, J. Mol. Biol. , 939 (1996).

[58] Tolman, J. Chem. Phys. **17**, 333 (1949).

[59] Salvino, Phys. Rev. B **34**, 6351 (1986).

[60] Sharp, Science **252**, 106 (1991).

[61] Sharp, Biochemistry **30**, 9686 (1991).

[62] Su, Ind. Eng. Chem. Res. **35**, 3399 (1996).

[63] Rashke, PNAS **98**, 5965.

[64] Markin, J. Phys. Chem B. **106**, 11810.

[65] Tsodikov, J. Comput. Chem. **23**, 600 (2002).

[66] Coleman, PROTEINS, Stru. func. bio. **61**, 1068.

[67] P. Ferrara, J. Apostolakis, and A. Caflisch., J. Phys. Chem. B **104**, 5000 (2000).

[68] S. A. Hassan, E. L. Mehler, D. Zhang, and H. Weinstein, Prot. Str. Fun. Bio. **51**, 109 (2003).

[69] Munoz, **390**, 196 (1997).

[70] Munoz, Proc. Natl. Acad. Sci. USA **95**, 5872 (1999).

[71] T. Lazaridis and M. Karplus, Prot. Str. Fun. Bio. **35**, 133 (1999).

[72] A. Vitalis and R. V. Pappu, J. Comput. Chem. **30**, 673 (2008).

[73] PDB codes: 1igd 1vii 2cyu 1crn 1enh 2ci2 2a3d 1ubq 1pht, (2009).

Andrea Prunotto Curriculum Vitæ (English)

PERSONAL DETAILS

Name:	Andrea
Surname:	Prunotto
Birth:	24.01.1975, Torino, Italy
Nationality:	Italian
Family status:	Single
Address:	Mutschellenstr. 89 CH-8038, Zürich
Contact:	Phone: 043 300 45 54 ; Mobile: 076 344 75 56
	Email: andrea.prunotto@gmail.com

WORK HISTORY

- **June 2005 - June 2009** Doktorand-Assistent at UNI-Zürich (computational structural biochemistry: design and development of an implicit solvent model by comparison between computer simulations results and NMR-X-ray experimental data); Prof. A. Caflisch, Biochemistry Institute, Winterthurerstr. 190, CH-8057, Zürich; Phone: +41 44 645 55 21 - Fax: +41 44 653 68 62; www.biochem-caflisch.unizh.ch
- **October 2004 - March 2005** Mathematics tutor for autistic people at Pegaso Centro Studi & Formazione SNC, Via Santa Croce 44 - 10024 Moncalieri (Torino); Phone: +39 011 64 38 37 - Fax: +39 011 64 87 604; www.pegasocsf.com
- **January 2004- June 2004** Teacher of Mathematics & Natural Science (full time) at Istituto Scolastico Paritario Francesco Faà di Bruno, Via San Donato, 31 - 10144 (Torino); Phone and Fax: +39 011 48 91 47; www.faadibruno.net
- **1994-2005** Music teacher and organist (part time) at Gesù Nazareno Parish, Via Pietro Palmieri 39 - 10138 Torino. Private lessons in mathematics and physics

EDUCATION

- **April 2010** ([1]) PhD Degree in Biochemistry, Universität Zürich
- **April 2004** Master Degree (Laurea) in Physics (110/110), Università di Torino
- **June 2001** Diploma in Music Composing (5th year), Conservatorio "G. Verdi", Torino
- **June 1997** Degree in Music theory and *solfeggio*, Conservatorio "G. Verdi", Via Mazzini, 11 - 10123 Torino; Phone +39 011 88 84 70 / +39 011 81 78 458; Fax: +39 011885165; www.conservatoriotorino.eu
- **July 1994** High School Diploma, ITIS "A. Avogadro", corso S. Maurizio, 8 - 10124 Torino; Phone: +39 011 81 53 611; Fax: +39 011 81 53 700; www.itisavogadro.it

[1] Scheduled

SKILLS

- **Languages:** English (intermediate), French (scholastic), German (beginner), Italian (mother tongue)
- **Programming:** C++, Python, Fortran, HTML, AWK, bash
- **Computer Science:** Linux: LaTeX, Xfig, Xmgrace, Gnuplot, VRML, Gnumeric; Windows: Word, Excel, Powerpoint
- **Music:** Composing, piano, guitar

OTHERS

- **November 2008** Translation (into Italian) of *Chaos: a very short introduction*, ISBN10: 019 285 3783, by L. Smith (Oxford University Press), for Codice Edizioni (*Caos*, ISBN10 978 88 7578 113 2, codiceedizioni.it/catalogo/pubblicazioni/caos)
- **September 2007** Review of *I fulmini Globulari* (*Ball lightnings*), ISBN10: 8875 076 960, by A. Carbognani (Macro Edizioni), for Codice Edizioni SRL, Via Giuseppe Pomba 17 - 10123 (Torino); Tel. +39 011 19 70 05 79/80 - Fax +39 011 19 70 05 82; www.codiceedizioni.it
- **December 2004 - April 2005** Theatrical perfomance *Il Teatro della Scienza*, in collaboration with Centro Scienza and La Stampa: *19.12.2004* at Teatro Colosseo, Via Madama Cristina, 71 - 10125 Torino; Phone: 011 66 98 034; www.teatrocolosseo.it; *18.04.2005* at Museo Tridentino di Scienze Naturali, Via Calepina, 14 - 38100 Trento; Phone: +39 046 127 03 11 - Fax: +39 046 123 38 30; www.mtsn.tn.it

Andrea Prunotto Curriculum Vitæ (Deutsch)

PERSÖNLICHEN DATEN

Vorname und Name:	Andrea Prunotto
Geburtsdatum:	24.01.1975
Geburtsort:	Torino, Italien
Nationalität:	Italien
Zivilstand:	Ledig
Adresse:	Mutschellenstr. 89 CH-8038, Zürich
Kontakt:	Tel.: 043 300 45 54 ; Nat.: 076 344 75 56
	Email: andrea.prunotto@gmail.com

BERUFLICHE TÄTIGKEITEN

- **Juni 2005 - Juni 2009** Doktorand-Assistent, UNI-Zürich (Projekt und Entwicklung eines "Implicit Solvent Model" im Vergleich zwischen MD Simulationen und NMR-X-ray Daten); Prof. A. Caflisch, Biochemisches Institut, Winterthurerstr. 190, CH-8057, Zürich; Tel.: +41 44 645 55 21 - Fax: +41 44 653 68 62; www.biochem-caflisch.unizh.ch
- **Oktober 2004 - März 2005** Mathematik Betreuer für autistische Personen, Pegaso Centro Studi & Formazione SNC, Via Santa Croce 44 - 10024 Moncalieri (Torino); Tel.: +39 011 64 38 37 - Fax: +39 011 64 87 604; www.pegasocsf.com
- **Januar 2004- Juni 2004** Mathematik - und Naturwissenschaft Lehrer (vollzeit), Istituto Scolastico Paritario Francesco Faà di Bruno, Via San Donato, 31 - 10144 (Torino); Tel. und Fax: +39 011 48 91 47; www.faadibruno.net
- **1994-2005** Musiklehrer und Organist (teilzeit), Gesù Nazareno Pfarrei, Via Pietro Palmieri 39 - 10138 Torino. Privatunterricht in Mathematik und Physik

SCHULE

- **April 2010** ([2]) Dr. Phil. in Biochemie, Universität Zürich
- **April 2004** Magister (Laurea) in Physik (110/110), Università di Torino
- **Juni 2001** Diplom in Komposition (5 jahr), Conservatorio "G. Verdi", Torino
- **Juni 1997** Diplom in *Solfeggio*, Conservatorio "G. Verdi", Via Mazzini, 11 - 10123 Torino; Tel. +39 011 88 84 70 / +39 011 81 78 458; Fax: +39 011885165;

[2] Angesetzt

- **Juli 1994** Matura (Informatik), ITIS "A. Avogadro", corso S. Maurizio, 8 - 10124 Torino; Tel.: +39 011 81 53 611; Fax: +39 011 81 53 700; www.itisavogadro.it

KENNTISSE

- **Sprachen:** Englisch (intermediate), Französich (mündlich), Deutsch (Anfänger), Italienisch (Muttersprache)
- **Programming:** C++, Python, Fortran, HTML, AWK, bash
- **Computer:** Linux: LaTeX, Xfig, Xmgrace, Gnuplot, VRML, Gnumeric; Windows: Word, Excel, Powerpoint
- **Musik:** Klavier, Gitarre, Komposition

WEITERE ERFAHRUNGEN

- **November 2008** Umsetzung (auf Italienisch): *Chaos: a very short introduction*, ISBN10: 019 285 3783, by L. Smith (Oxford University Press), für Codice Edizioni (*Caos*, ISBN10 978 88 7578 113 2, codiceedizioni.it/catalogo/pubblicazioni/caos)
- **September 2007** Review: *I fulmini Globulari* (*Die Kugelblitzen*), ISBN10: 8875 076 960, by A. Carbognani (Macro Edizioni), für Codice Edizioni SRL, Via Giuseppe Pomba 17 - 10123 (Torino); Tel. +39 011 19 70 05 79/80 - Fax +39 011 19 70 05 82; www.codiceedizioni.it
- **Dezember 2004 - April 2005** Theatervorstellung: *Il Teatro della Scienza*, für Centro Scienza und La Stampa *19.12.2004* Teatro Colosseo, Via Madama Cristina, 71 - 10125 Torino; Tel.: 011 66 98 034; www.teatrocolosseo.it; *18.04.2005* Museo Tridentino di Scienze Naturali, Via Calepina, 14 - 38100 Trento; Tel.: +39 046 127 03 11 - Fax: +39 046 123 38 30; www.mtsn.tn.it